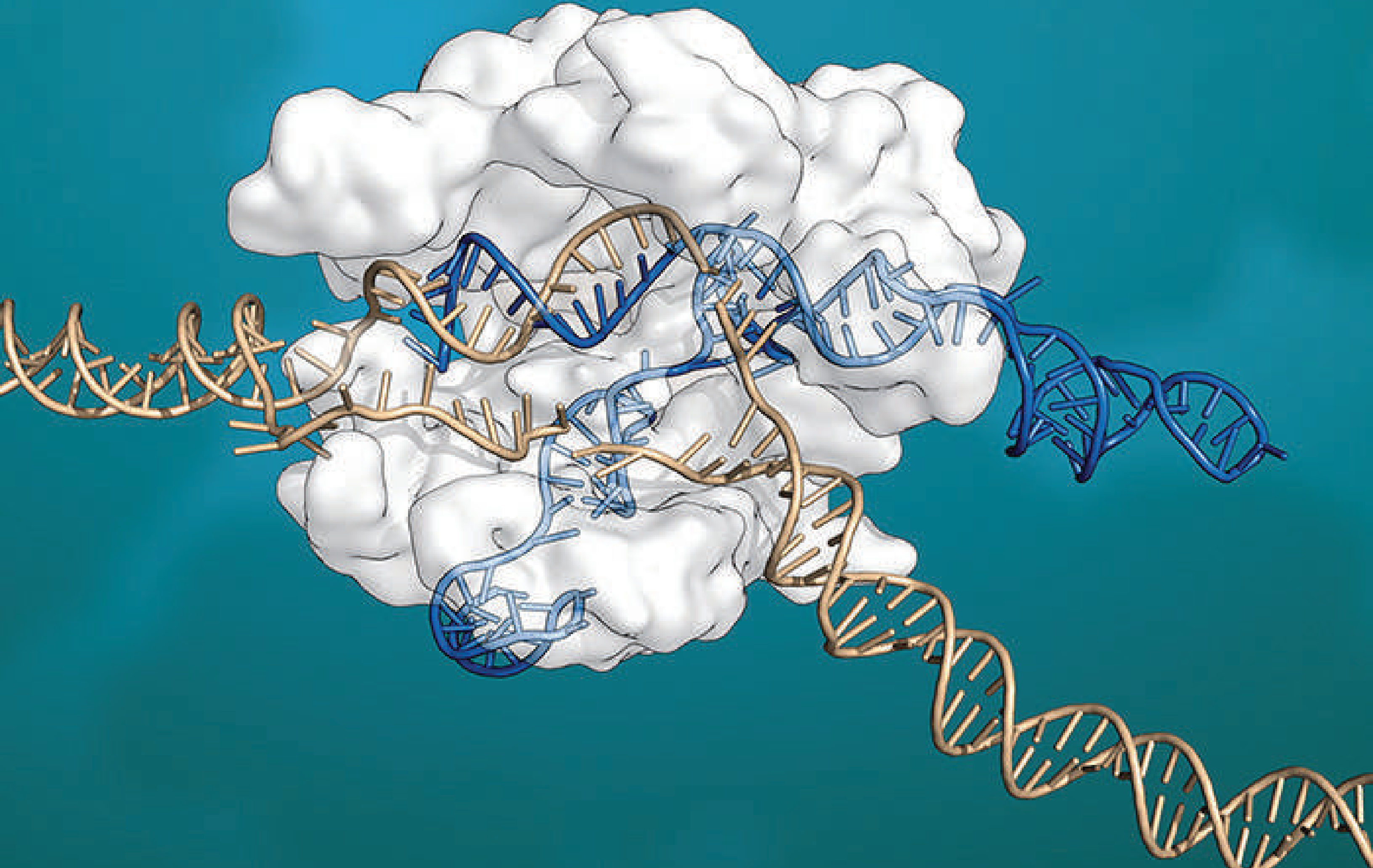
CONCEPTS OF

GENETICS

TWELFTH EDITION



Klug | Cummings | Spencer

Palladino | Killian

Brief Contents

PART ONE

GENES, CHROMOSOMES, AND HEREDITY

PART TWO

DNA: STRUCTURE, REPLICATION, AND ORGANIZATION

PART THREE

GENE EXPRESSION AND ITS REGULATION

PART FOUR

GENETIC TECHNOLOGY AND GENOMICS

PART FIVE

GENETIC ANALYSIS OF ORGANISMS AND POPULATIONS

SPECIAL TOPICS IN MODERN GENETICS

Nobel Prizes Awarded for Research in Genetics or Genetics-Related Areas

Year	Recipients	Nobel Prize	Discovery/Research Topic
2017	J. C. Hall M. Rosbash M. W. Young	Physiology or Medicine	Identification of the genes and molecular mechanisms that regulate circadian rhythms
2015	T. Lindahl P. Modrich A. Sancar	Chemistry	Mechanistic studies of DNA repair
2012	J. B. Gurdon S. Yamanaka	Physiology or Medicine	Differentiated cells can be reprogrammed to become pluripotent
2009	V. Ramakrishnan T. A. Steitz A. E. Yonath	Chemistry	Structure and function of the ribosome
2009	E. H. Blackburn C. W. Greider J. W. Szostak	Physiology or Medicine	The nature and replication of the DNA of telomeres, and the discovery of the telomere-replenishing ribonucleoprotein enzyme telomerase
2008	O. Shimomura M. Chalfie R. Y. Tsien	Chemistry	Discovery and development of a genetically encoded fluorescent protein as an *in vivo* marker of gene expression
2007	M. R. Capecchi M. J. Evans O. Smithies	Physiology or Medicine	Gene-targeting technology essential to the creation of knockout mice serving as animal models of human disease
2006	R. D. Kornberg	Chemistry	Molecular basis of eukaryotic transcription
2006	A. Z. Fire C. C. Mello	Physiology or Medicine	Gene silencing using RNA interference (RNAi)
2004	A. Ciechanover A. Hershko I. Rose	Chemistry	Regulation of protein degradation by the proteasome
2002	S. Brenner H. R. Horvitz J. E. Sulston	Physiology or Medicine	Genetic regulation of organ development and programmed cell death (apoptosis)
2001	L. H. Hartwell T. Hunt P. M. Nurse	Physiology or Medicine	Genes and regulatory molecules controlling the cell cycle
1999	G. Blobel	Physiology or Medicine	Genetically encoded amino acid sequences in proteins that guide their cellular transport
1997	S. B. Prusiner	Physiology or Medicine	Prions—a new biological principle of infection
1995	E. B. Lewis C. Nüsslein-Volhard E. Wieschaus	Physiology or Medicine	Genetic control of early development in *Drosophila*
1993	R. J. Roberts P. A. Sharp	Physiology or Medicine	RNA processing of split genes
	K. B. Mullis M. Smith	Chemistry	Development of polymerase chain reaction (PCR) and site-directed mutagenesis (SDM)
1989	J. M. Bishop H. E. Varmus	Physiology or Medicine	Role of retroviruses and oncogenes in cancer
	T. R. Cech S. Altman	Chemistry	Ribozyme function during RNA splicing
1987	S. Tonegawa	Physiology or Medicine	Genetic basis of antibody diversity

Year	Recipients	Nobel Prize	Discovery/Research Topic
1985	M. S. Brown J. L. Goldstein	Physiology or Medicine	Genetic regulation of cholesterol metabolism
1983	B. McClintock	Physiology or Medicine	Mobile genetic elements in maize
1982	A. Klug	Chemistry	Crystalline structure analysis of significant complexes, including tRNA and nucleosomes
1980	P. Berg W. Gilbert F. Sanger	Chemistry	Development of recombinant DNA and DNA sequencing technology
1978	W. Arber D. Nathans H. O. Smith	Physiology or Medicine	Recombinant DNA technology using restriction endonuclease technology
1976	B. S. Blumberg D. C. Gajdusek	Physiology or Medicine	Elucidation of the human prion-based diseases, kuru and Creutzfeldt-Jakob disease
1975	D. Baltimore R. Dulbecco H. M. Temin	Physiology or Medicine	Molecular genetics of tumor viruses
1972	G. M. Edelman R. R. Porter	Physiology or Medicine	Chemical structure of immunoglobulins
	C. B. Anfinsen	Chemistry	Relationship between primary and tertiary structure of proteins
1970	N. Borlaug	Peace Prize	Genetic improvement of Mexican wheat
1969	M. Delbrück A. D. Hershey S. E. Luria	Physiology or Medicine	Replication mechanisms and genetic structure of bacteriophages
1968	H. G. Khorana M. W. Nirenberg R. W. Holley	Physiology or Medicine	For their interpretation of the genetic code and its function during protein synthesis
1966	P. F. Rous	Physiology or Medicine	Viral induction of cancer in chickens
1965	F. Jacob A. M. Lwoff J. L. Monod	Physiology or Medicine	Genetic regulation of enzyme synthesis in bacteria
1962	F. H. C. Crick J. D. Watson M. H. F. Wilkins	Physiology or Medicine	Double helical model of DNA
	J. C. Kendrew M. F. Perutz	Chemistry	Three-dimensional structure of globular proteins
1959	A. Kornberg S. Ochoa	Physiology or Medicine	Biological synthesis of DNA and RNA
1958	G. W. Beadle E. L. Tatum	Physiology or Medicine	Genetic control of biochemical processes
	J. Lederberg	Physiology or Medicine	Genetic recombination in bacteria
	F. Sanger	Chemistry	Primary structure of proteins
1954	L. C. Pauling	Chemistry	Alpha helical structure of proteins
1946	H. J. Müller	Physiology or Medicine	X-ray induction of mutations in *Drosophila*
1933	T. H. Morgan	Physiology or Medicine	Chromosomal theory of inheritance
1930	K. Landsteiner	Physiology or Medicine	Discovery of human blood groups

CONCEPTS OF GENETICS

This page intentionally left blank

CONCEPTS OF GENETICS

TWELFTH EDITION

William S. Klug
THE COLLEGE OF NEW JERSEY

Michael R. Cummings
ILLINOIS INSTITUTE OF TECHNOLOGY

Charlotte A. Spencer
UNIVERSITY OF ALBERTA

Michael A. Palladino
MONMOUTH UNIVERSITY

Darrell J. Killian
COLORADO COLLEGE

Pearson

330 Hudson Street, NY NY 10013

Library of Congress Cataloging-in-Publication Data
Names: Klug, William S., author. | Cummings, Michael R., author. | Spencer, Charlotte A., author. | Palladino, Michael A., author. | Killian, Darrell J., author.
Title: Concepts of genetics / William S. Klug (The College of New Jersey), Michael R. Cummings (Illinois Institute of Technology), Charlotte A. Spencer, Michael A. Palladino (Monmouth University), Darrell J. Killian (Colorado College).
Description: Twelfth edition. | Hoboken, New Jersey : Pearson Education, Inc., 2018. | Includes index.
Identifiers: LCCN 2017047484 | ISBN 9780134604718
Subjects: LCSH: Genetics—Textbooks.
Classification: LCC QH430 .K574 2018 | DDC 572.8—dc23
LC record available at https://lccn.loc.gov/2017047484

Courseware Portfolio Manager: Michael Gillespie
Content Producer: Brett Coker
Managing Producer: Mike Early
Product Marketing Manager: Christa Pesek Pelaez
Courseware Director, Content Development: Ginnie Simione Jutson
Courseware Senior Analysts, Content Development: Sonia Divittorio and Barbara Price
Courseware Portfolio Management Assistant: Summer Giles
Rich Media Content Producer: Chloe Veylit
Full-Service Vendor: SPi Global
Full-Service Project Managers: Michelle Gardner and Heidi Aguiar
Copyeditor: Lucy Mullins
Illustrations: Imagineering Art
Art Coordinator: Stephanie Marquez
Design Manager: Maria Guglielmo
Cover and Interior Designer: Tamara Newnam
Rights & Permissions Project Manager: Angelica Aranas
Rights & Permissions Management: Eric Schrader
Photo Researcher: Eric Schrader
Manufacturing Buyer: Stacey J. Weinberger

Cover Illustration: H. Adam Steinberg, artforscience.com

Pearson

ISBN 10: 0-134-60471-7; ISBN 13: 978-0-134-60471-8 (Student edition)
ISBN 10: 0-134-81892-X; ISBN 13: 978-0-134-81892-4 (Books à la Carte edition)
1 18
www.pearson.com

About the Authors

William S. Klug is an Emeritus Professor of Biology at The College of New Jersey (formerly Trenton State College) in Ewing, New Jersey, where he served as Chair of the Biology Department for 17 years. He received his B.A. degree in Biology from Wabash College in Crawfordsville, Indiana, and his Ph.D. from Northwestern University in Evanston, Illinois. Prior to coming to The College of New Jersey, he was on the faculty of Wabash College, where he first taught genetics, as well as general biology and electron microscopy. His research interests have involved ultrastructural and molecular genetic studies of development, utilizing oogenesis in *Drosophila* as a model system. He has taught the genetics course as well as the senior capstone seminar course in Human and Molecular Genetics to undergraduate biology majors for over four decades. He was the recipient in 2001 of the first annual teaching award given at The College of New Jersey, granted to the faculty member who "most challenges students to achieve high standards." He also received the 2004 Outstanding Professor Award from Sigma Pi International, and in the same year, he was nominated as the Educator of the Year, an award given by the Research and Development Council of New Jersey. When not revising one of his textbooks, immersed in the literature of genetics, or trying to avoid double bogies, Dr. Klug can sometimes be found paddling in the Gulf of Mexico or in Maine's Penobscot Bay.

Michael R. Cummings is a Research Professor in the Department of Biological, Chemical, and Physical Sciences at Illinois Institute of Technology, Chicago, Illinois. For more than 25 years, he was a faculty member in the Department of Biological Sciences and in the Department of Molecular Genetics at the University of Illinois at Chicago. He has also served on the faculties of Northwestern University and Florida State University. He received his B.A. from St. Mary's College in Winona, Minnesota, and his M.S. and Ph.D. from Northwestern University in Evanston, Illinois. In addition to this text, he has written textbooks in human genetics and general biology. His research interests center on the molecular organization and physical mapping of the heterochromatic regions of human acrocentric chromosomes. At the undergraduate level, he teaches courses in molecular genetics, human genetics, and general biology, and has received numerous awards for teaching excellence given by university faculty, student organizations, and graduating seniors. When not teaching or writing, Dr. Cummings can often be found far offshore fishing for the one that got away.

Charlotte A. Spencer is a retired Associate Professor from the Department of Oncology at the University of Alberta in Edmonton, Alberta, Canada. She has also served as a faculty member in the Department of Biochemistry at the University of Alberta. She received her B.Sc. in Microbiology from the University of British Columbia and her Ph.D. in Genetics from the University of Alberta, followed by postdoctoral training at the Fred Hutchinson Cancer Research Center in Seattle, Washington. Her research interests involve the regulation of RNA polymerase II transcription in cancer cells, cells infected with DNA viruses, and cells traversing the mitotic phase of the cell cycle. She has taught undergraduate and graduate courses in biochemistry, genetics, molecular biology, and oncology. She has also written booklets in the Prentice Hall Exploring Biology series. When not writing and editing contributions to genetics textbooks, Dr. Spencer works on her hazelnut farm and enjoys the peace and quiet of a remote Island off the west coast of British Columbia.

Michael A. Palladino is Vice Provost for Graduate Studies, former Dean of the School of Science, and Professor of Biology at Monmouth University in West Long Branch, New Jersey. He received his B.S. degree in Biology from The College of New Jersey and his Ph.D. in Anatomy and Cell Biology from the University of Virginia. For more than 15 years he directed a laboratory of undergraduate student researchers supported by external funding from the National Institutes of Health, biopharma companies, and other agencies. He and his undergraduates studied molecular mechanisms involved in innate immunity of mammalian male reproductive organs and genes involved in oxygen homeostasis and ischemic injury of the testis. He has taught a wide range of courses including genetics, biotechnology, endocrinology, and cell and molecular biology. He has received several awards for research and teaching, including the 2009 Young Andrologist Award of the American Society of Andrology, the 2005 Distinguished Teacher Award from Monmouth University, and the 2005 Caring Heart Award from the New Jersey Association for Biomedical Research. He is co-author of the undergraduate textbook *Introduction to Biotechnology*. He was Series Editor for the Benjamin Cummings *Special Topics in Biology* booklet series, and author of the first booklet in the series, *Understanding the Human Genome Project.* When away from the university or authoring textbooks, Dr. Palladino can often be found watching or playing soccer or attempting to catch most any species of fish in freshwater or saltwater.

Darrell J. Killian is an Associate Professor and current Chair of the Department of Molecular Biology at Colorado College in Colorado Springs, Colorado. He received his B.A. degree in Molecular Biology and Biochemistry from Wesleyan University in Middletown, Connecticut, prior to working as a Research Technician in Molecular Genetics at Rockefeller University in New York, New York. He earned his Ph.D. in Developmental Genetics from New York University in New York, New York, and received his postdoctoral training at the University of Colorado—Boulder in the Department of Molecular, Cellular, and Developmental Biology. Prior to joining Colorado College, he was an Assistant Professor of Biology at the College of New Jersey in Ewing, New Jersey. His research focuses on the genetic regulation of animal development, and he has received funding from the National Institutes of Health and the National Science Foundation. Currently, he and his undergraduate research assistants are investigating the molecular genetic regulation of nervous system development using *C. elegans* and *Drosophila* as model systems. He teaches undergraduate courses in genetics, molecular and cellular biology, stem cell biology, and developmental neurobiology. When away from the classroom and research lab, Dr. Killian can often be found on two wheels exploring trails in the Pike and San Isabel National Forests.

Dedication

We dedicate this edition to our long-time colleague and friend Harry Nickla, who sadly passed away in 2017. With decades of experience teaching Genetics to students at Creighton University, Harry's contribution to our texts included authorship of the *Student Handbook and Solutions Manual* and the test bank, as well as devising most of the Extra Spicy problems at the end of each chapter. He was also a source of advice during the planning session for each new edition, and during our many revisions. We always appreciated his professional insights, friendship, and conviviality. We were lucky to have him as part of our team, and we miss him greatly.

WSK, MRC, CAS, MAP, and DJK

Brief Contents

PART ONE
GENES, CHROMOSOMES, AND HEREDITY

PART TWO
DNA: STRUCTURE, REPLICATION, AND ORGANIZATION

PART THREE
GENE EXPRESSION AND ITS REGULATION

PART FOUR
GENETIC TECHNOLOGY AND GENOMICS

PART FIVE
GENETIC ANALYSIS OF ORGANISMS AND POPULATIONS

SPECIAL TOPICS IN MODERN GENETICS

This page intentionally left blank

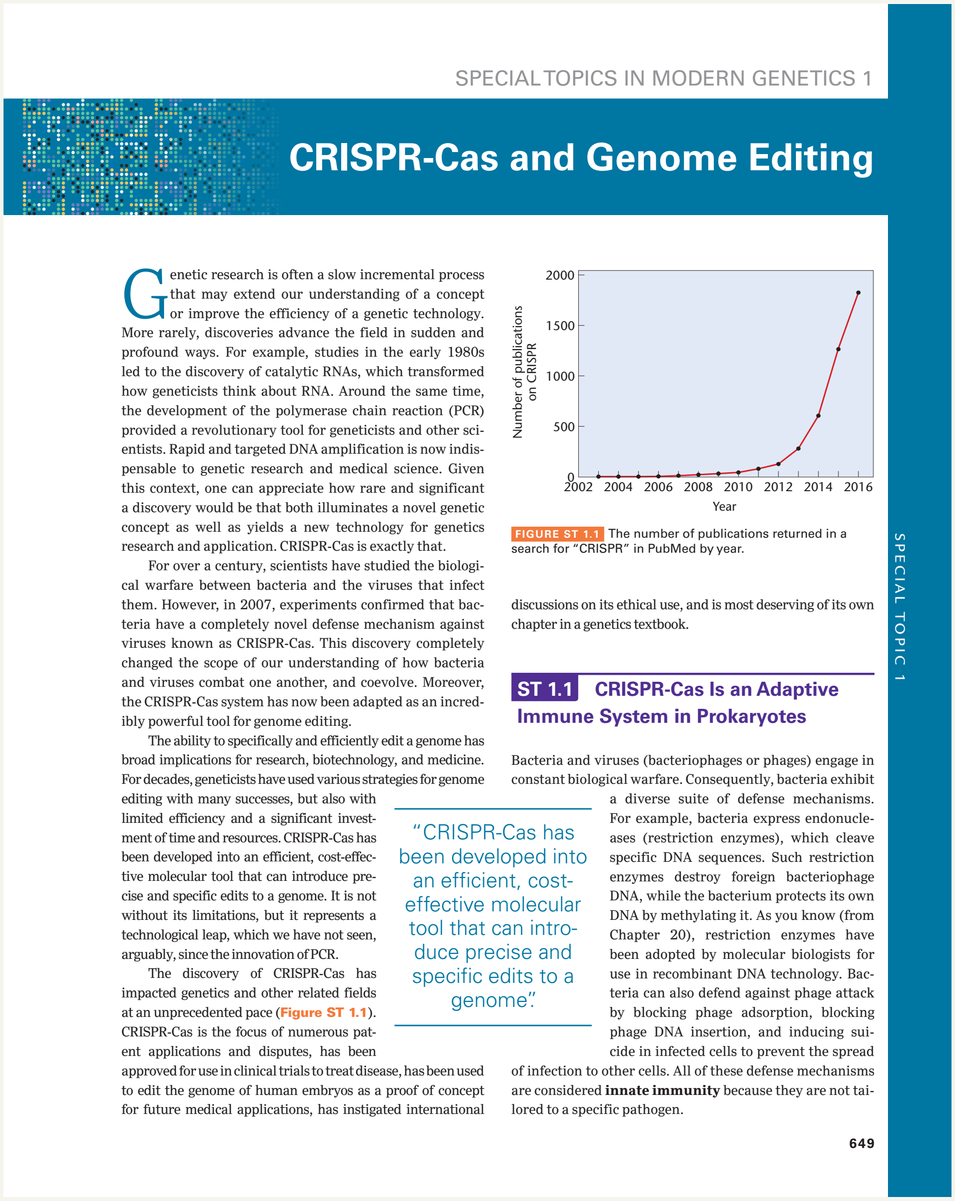
Explore Cutting-Edge Topics

Concepts of Genetics **emphasizes the fundamental ideas of genetics, while exploring modern techniques and applications of genetic analysis. This best-selling text continues to provide understandable explanations of complex, analytical topics and recognizes the importance of teaching students how to become effective problem solvers.**

Six **Special Topics in Modern Genetics** mini-chapters concisely explore cutting-edge, engaging, and relevant topics.

- **NEW!** CRISPR-Cas and Genome Editing
- DNA Forensics
- Genomics and Precision Medicine
- Genetically Modified Foods
- Gene Therapy
- **NEW!** Advances in Neurogenetics: The Study of Huntington Disease

Special Topic chapters include Review and Discussion questions, which are also assignable in Mastering Genetics.



SPECIAL TOPICS IN MODERN GENETICS 1

CRISPR-Cas and Genome Editing

Genetic research is often a slow incremental process that may extend our understanding of a concept or improve the efficiency of a genetic technology. More rarely, discoveries advance the field in sudden and profound ways. For example, studies in the early 1980s led to the discovery of catalytic RNAs, which transformed how geneticists think about RNA. Around the same time, the development of the polymerase chain reaction (PCR) provided a revolutionary tool for geneticists and other scientists. Rapid and targeted DNA amplification is now indispensable to genetic research and medical science. Given this context, one can appreciate how rare and significant a discovery would be that both illuminates a novel genetic concept as well as yields a new technology for genetics research and application. CRISPR-Cas is exactly that.

For over a century, scientists have studied the biological warfare between bacteria and the viruses that infect them. However, in 2007, experiments confirmed that bacteria have a completely novel defense mechanism against viruses known as CRISPR-Cas. This discovery completely changed the scope of our understanding of how bacteria and viruses combat one another, and coevolve. Moreover, the CRISPR-Cas system has now been adapted as an incredibly powerful tool for genome editing.

The ability to specifically and efficiently edit a genome has broad implications for research, biotechnology, and medicine. For decades, geneticists have used various strategies for genome editing with many successes, but also with limited efficiency and a significant investment of time and resources. CRISPR-Cas has been developed into an efficient, cost-effective molecular tool that can introduce precise and specific edits to a genome. It is not without its limitations, but it represents a technological leap, which we have not seen, arguably, since the innovation of PCR.

"CRISPR-Cas has been developed into an efficient, cost-effective molecular tool that can introduce precise and specific edits to a genome".

The discovery of CRISPR-Cas has impacted genetics and other related fields at an unprecedented pace (**Figure ST 1.1**). CRISPR-Cas is the focus of numerous patent applications and disputes, has been approved for use in clinical trials to treat disease, has been used to edit the genome of human embryos as a proof of concept for future medical applications, has instigated international discussions on its ethical use, and is most deserving of its own chapter in a genetics textbook.

FIGURE ST 1.1 The number of publications returned in a search for "CRISPR" in PubMed by year.

ST 1.1 CRISPR-Cas Is an Adaptive Immune System in Prokaryotes

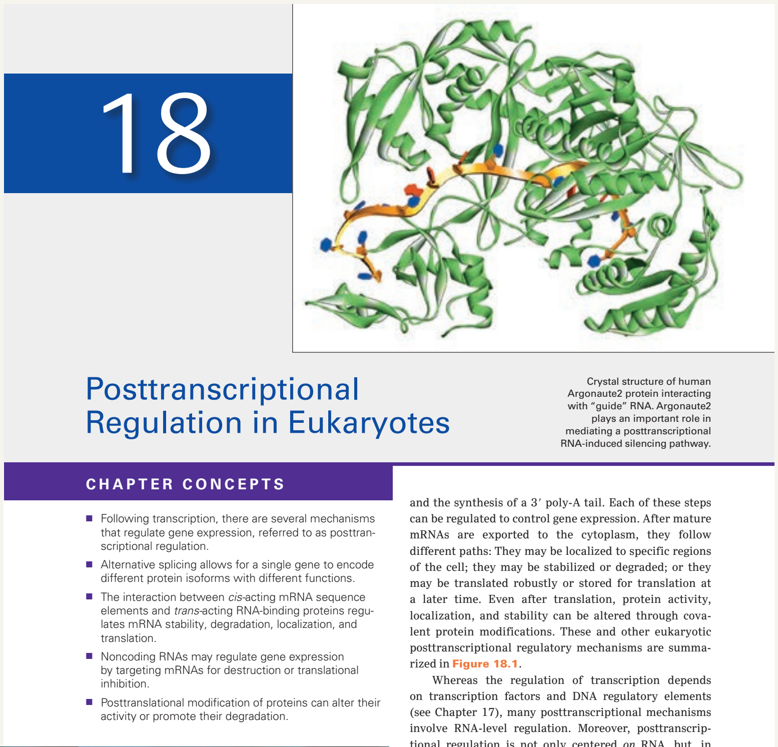
Bacteria and viruses (bacteriophages or phages) engage in constant biological warfare. Consequently, bacteria exhibit a diverse suite of defense mechanisms. For example, bacteria express endonucleases (restriction enzymes), which cleave specific DNA sequences. Such restriction enzymes destroy foreign bacteriophage DNA, while the bacterium protects its own DNA by methylating it. As you know (from Chapter 20), restriction enzymes have been adopted by molecular biologists for use in recombinant DNA technology. Bacteria can also defend against phage attack by blocking phage adsorption, blocking phage DNA insertion, and inducing suicide in infected cells to prevent the spread of infection to other cells. All of these defense mechanisms are considered **innate immunity** because they are not tailored to a specific pathogen.

649

Explore the Latest Updates

The 12th edition has been heavily updated throughout, including a reorganization and expansion of coverage of gene regulation in eukaryotes. This expansion reflects our growing knowledge of the critical roles RNA and epigenetics play in regulating gene activity.

NEW! Gene regulation in eukaryotes has been expanded into three chapters: transcriptional regulation (Ch. 17), posttranscriptional regulation (Ch. 18), and epigenetic regulation (Ch. 19).



18

Posttranscriptional Regulation in Eukaryotes

Crystal structure of human Argonaute2 protein interacting with "guide" RNA. Argonaute2 plays an important role in mediating a posttranscriptional RNA-induced silencing pathway.

CHAPTER CONCEPTS

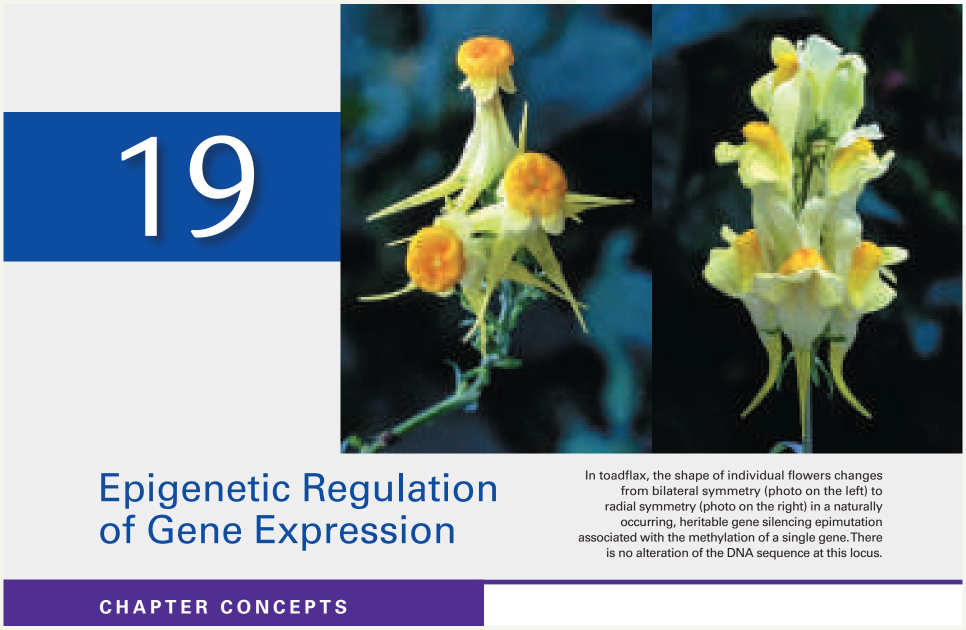
- Following transcription, there are several mechanisms that regulate gene expression, referred to as posttranscriptional regulation.
- Alternative splicing allows for a single gene to encode different protein isoforms with different functions.
- The interaction between *cis*-acting mRNA sequence elements and *trans*-acting RNA-binding proteins regulates mRNA stability, degradation, localization, and translation.
- Noncoding RNAs may regulate gene expression by targeting mRNAs for destruction or translational inhibition.
- Posttranslational modification of proteins can alter their activity or promote their degradation.

and the synthesis of a 3′ poly-A tail. Each of these steps can be regulated to control gene expression. After mature mRNAs are exported to the cytoplasm, they follow different paths: They may be localized to specific regions of the cell; they may be stabilized or degraded; or they may be translated robustly or stored for translation at a later time. Even after translation, protein activity, localization, and stability can be altered through covalent protein modifications. These and other eukaryotic posttranscriptional regulatory mechanisms are summarized in Figure 18.1.

Whereas the regulation of transcription depends on transcription factors and DNA regulatory elements (see Chapter 17), many posttranscriptional mechanisms involve RNA-level regulation. Moreover, posttranscriptional regulation is not only centered *on* RNA, but, in some cases, is regulated *by* RNA. Noncoding RNAs play important roles in the regulation of eukaryotic gene expression.

In this chapter, we will explore several important mechanisms and themes of eukaryotic posttranscriptional regulation. As you read on, keep in mind that while scientists have learned a great deal about how genes are regulated at the posttranscriptional level, there are still many unanswered questions for the curious student to ponder.

413



19

Epigenetic Regulation of Gene Expression

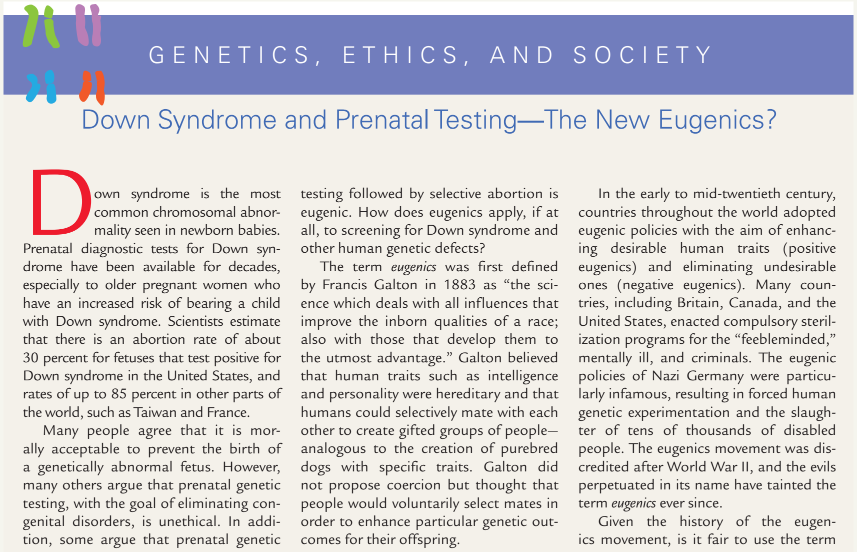
In toadflax, the shape of individual flowers changes from bilateral symmetry (photo on the left) to radial symmetry (photo on the right) in a naturally occurring, heritable gene silencing epimutation associated with the methylation of a single gene. There is no alteration of the DNA sequence at this locus.

CHAPTER CONCEPTS

NEW! A new chapter focuses on epigenetics, updating and expanding coverage that used to be in a Special Topics chapter.

and Ethical Considerations

With the rapid growth of our understanding of genetics and the ongoing introduction of powerful tools that can edit genes and genomes, it's important to encourage students to confront ethical issues and consider questions that arise in the study of genetics.



GENETICS, ETHICS, AND SOCIETY

Down Syndrome and Prenatal Testing—The New Eugenics?

Down syndrome is the most common chromosomal abnormality seen in newborn babies. Prenatal diagnostic tests for Down syndrome have been available for decades, especially to older pregnant women who have an increased risk of bearing a child with Down syndrome. Scientists estimate that there is an abortion rate of about 30 percent for fetuses that test positive for Down syndrome in the United States, and rates of up to 85 percent in other parts of the world, such as Taiwan and France.

Many people agree that it is morally acceptable to prevent the birth of a genetically abnormal fetus. However, many others argue that prenatal genetic testing, with the goal of eliminating congenital disorders, is unethical. In addition, some argue that prenatal genetic testing followed by selective abortion is eugenic. How does eugenics apply, if at all, to screening for Down syndrome and other human genetic defects?

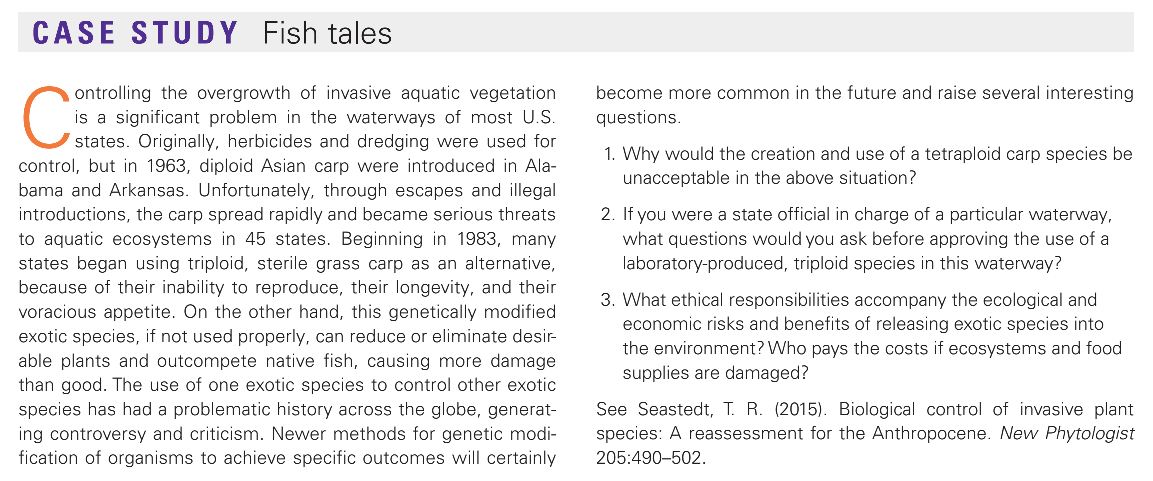
The term *eugenics* was first defined by Francis Galton in 1883 as "the science which deals with all influences that improve the inborn qualities of a race; also with those that develop them to the utmost advantage." Galton believed that human traits such as intelligence and personality were hereditary and that humans could selectively mate with each other to create gifted groups of people—analogous to the creation of purebred dogs with specific traits. Galton did not propose coercion but thought that people would voluntarily select mates in order to enhance particular genetic outcomes for their offspring.

In the early to mid-twentieth century, countries throughout the world adopted eugenic policies with the aim of enhancing desirable human traits (positive eugenics) and eliminating undesirable ones (negative eugenics). Many countries, including Britain, Canada, and the United States, enacted compulsory sterilization programs for the "feebleminded," mentally ill, and criminals. The eugenic policies of Nazi Germany were particularly infamous, resulting in forced human genetic experimentation and the slaughter of tens of thousands of disabled people. The eugenics movement was discredited after World War II, and the evils perpetuated in its name have tainted the term *eugenics* ever since.

Given the history of the eugenics movement, is it fair to use the term

NEW! Genetics, Ethics, and Society essays appear in many chapters. Each one provides a synopsis of an ethical issue, related to chapter content, that impacts society today. Each includes a section called Your Turn, directing students to resources to help them explore the issue and answer questions.

NEW and REVISED! Case Studies conclude each chapter, introducing a short vignette of an everyday genetics-related situation and posing several discussion questions, including one focusing on ethics.



CASE STUDY Fish tales

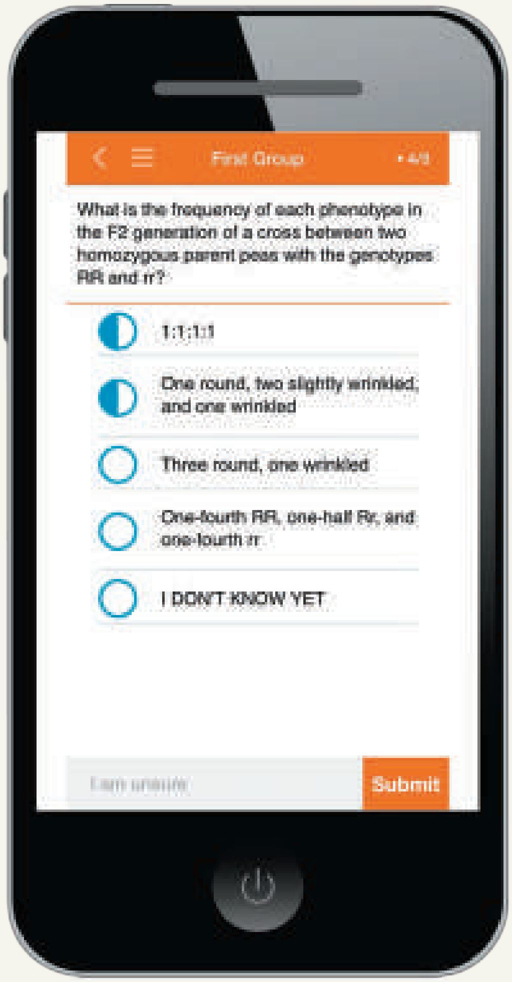
Controlling the overgrowth of invasive aquatic vegetation is a significant problem in the waterways of most U.S. states. Originally, herbicides and dredging were used for control, but in 1963, diploid Asian carp were introduced in Alabama and Arkansas. Unfortunately, through escapes and illegal introductions, the carp spread rapidly and became serious threats to aquatic ecosystems in 45 states. Beginning in 1983, many states began using triploid, sterile grass carp as an alternative, because of their inability to reproduce, their longevity, and their voracious appetite. On the other hand, this genetically modified exotic species, if not used properly, can reduce or eliminate desirable plants and outcompete native fish, causing more damage than good. The use of one exotic species to control other exotic species has had a problematic history across the globe, generating controversy and criticism. Newer methods for genetic modification of organisms to achieve specific outcomes will certainly become more common in the future and raise several interesting questions.

1. Why would the creation and use of a tetraploid carp species be unacceptable in the above situation?
2. If you were a state official in charge of a particular waterway, what questions would you ask before approving the use of a laboratory-produced, triploid species in this waterway?
3. What ethical responsibilities accompany the ecological and economic risks and benefits of releasing exotic species into the environment? Who pays the costs if ecosystems and food supplies are damaged?

See Seastedt, T. R. (2015). Biological control of invasive plant species: A reassessment for the Anthropocene. *New Phytologist* 205:490–502.

Learn Genetics Concepts and Problem Solving

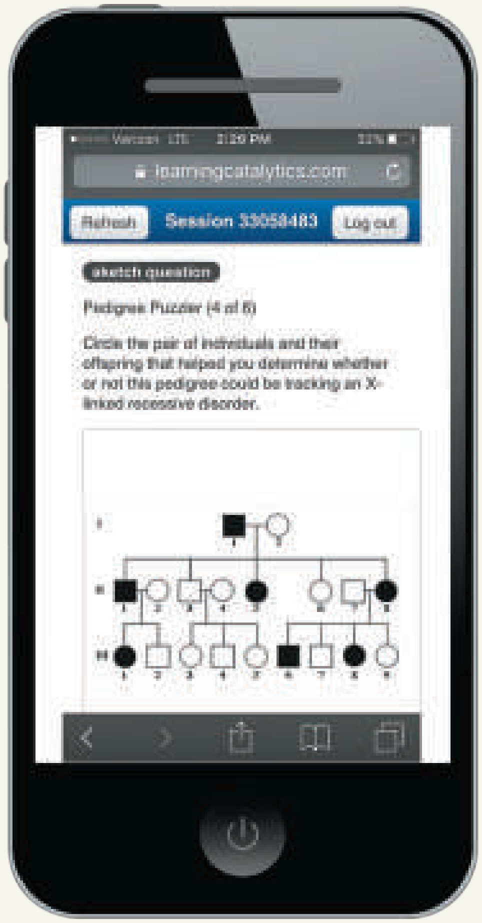
Mastering™ Genetics helps students master key genetics concepts while reinforcing problem-solving skills with hints and feedback specific to their misconceptions. Mastering Genetics includes content and tools for before, during, and after class. Learn more at www.pearson.com/mastering/genetics



What is the frequency of each phenotype in the F2 generation of a cross between two homozygous parent peas with the genotypes RR and rr?

- 1:1:1:1
- One round, two slightly wrinkled, and one wrinkled
- Three round, one wrinkled
- One-fourth RR, one-half Rr, and one-fourth rr
- I DON'T KNOW YET

NEW! Dynamic Study Modules personalize each student's learning experience. Available for assignments or for self-study, these chapter-based modules help prepare students for in-class discussions, problem solving, or active learning. A mobile app is available for iOS and Android devices.

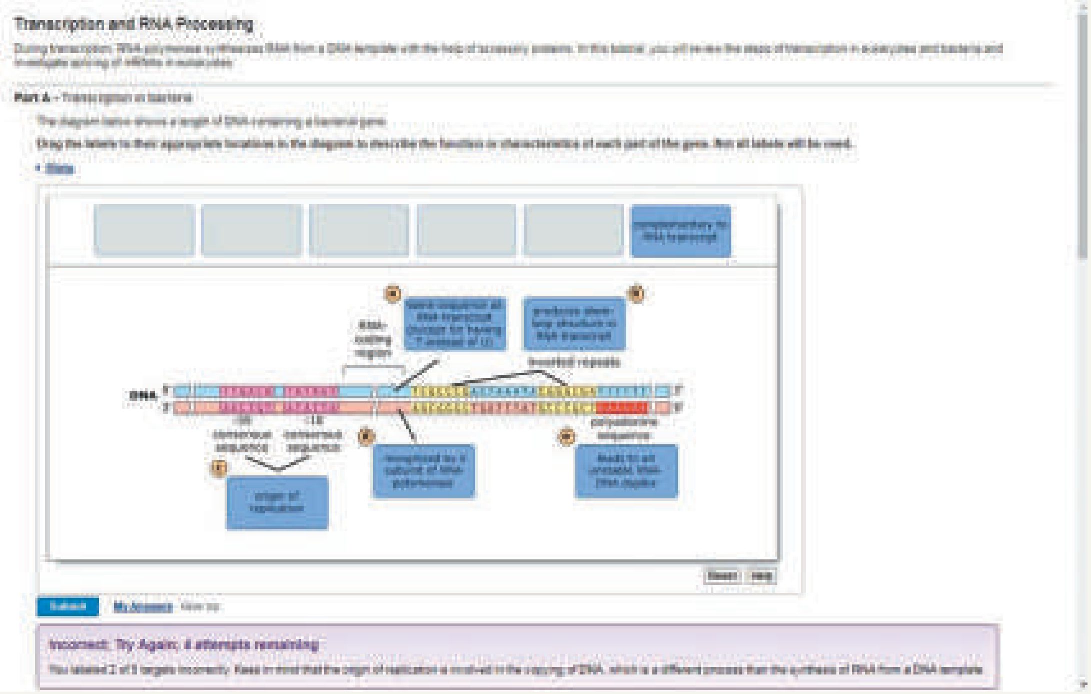


Pedigree Puzzler (4 of 6)

Circle the pair of individuals and their offspring that helped you determine whether or not this pedigree could be tracking an X-linked recessive disorder.

Learning Catalytics is a "bring your own device" (smartphone, tablet, or laptop) assessment and active classroom system that helps engage students. Instructors can create their own questions, draw from community content, or access Pearson's library of question clusters.

with Mastering Genetics



Transcription and RNA Processing

During transcription, RNA polymerase synthesizes RNA from a DNA template with the help of accessory proteins. In this tutorial, you will review the steps of transcription in eukaryotes and bacteria and investigate splicing of mRNAs in eukaryotes.

Part A - Transcription in bacteria

The diagram below shows a length of DNA containing a bacterial gene.

Drag the labels to their appropriate locations in the diagram to describe the function or characteristics of each part of the gene. Not all labels will be used.

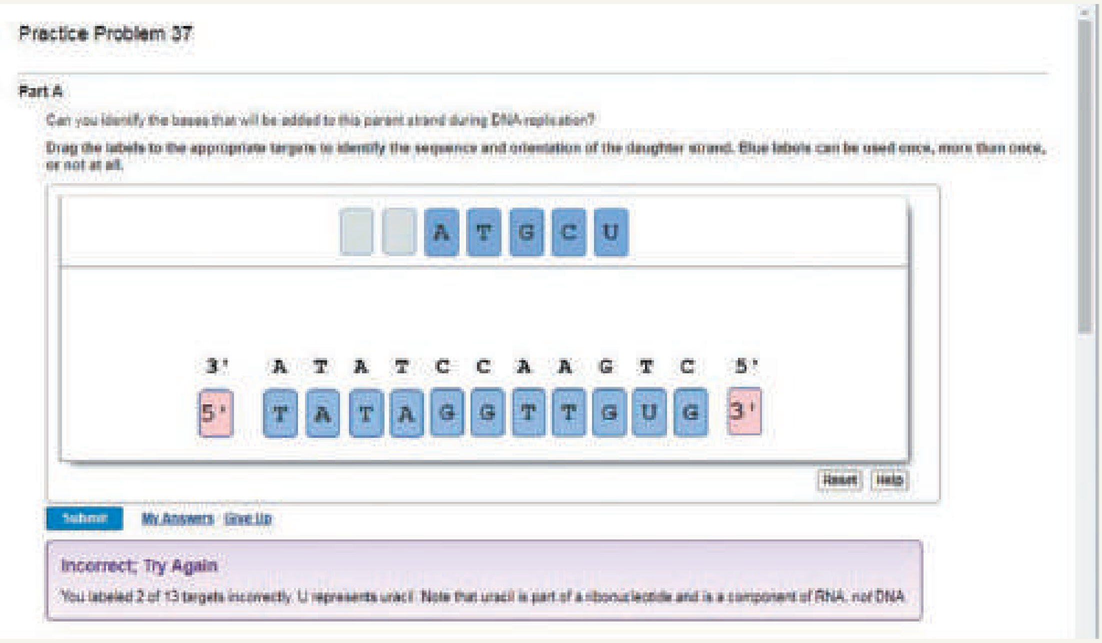
• Hints
complementary to RNA transcript
same sequence as RNA transcript (except for having T instead of U)
produces stem-loop structure in RNA transcript
RNA-coding region
inverted repeats
DNA
5'
3'
3'
5'
-35 consensus sequence
-10 consensus sequence
polypyrimidine sequence
origin of replication
recognized by σ subunit of RNA polymerase
leads to an unstable RNA-DNA duplex
Reset
Help
Submit
My Answers
Give Up

Incorrect; Try Again; 4 attempts remaining

You labeled 2 of 5 targets incorrectly. Keep in mind that the origin of replication is involved in the copying of DNA, which is a different process than the synthesis of RNA from a DNA template.

Tutorials and activities feature personalized wrong-answer feedback and hints that emulate the office-hour experience to guide student learning. New tutorials include coverage of topics like CRISPR-Cas.

100 Practice Problems offer more opportunities to develop problem-solving skills. These questions appear only in Mastering Genetics and include targeted wrong-answer feedback to help students learn.



Practice Problem 37

Part A

Can you identify the bases that will be added to this parent strand during DNA replication?

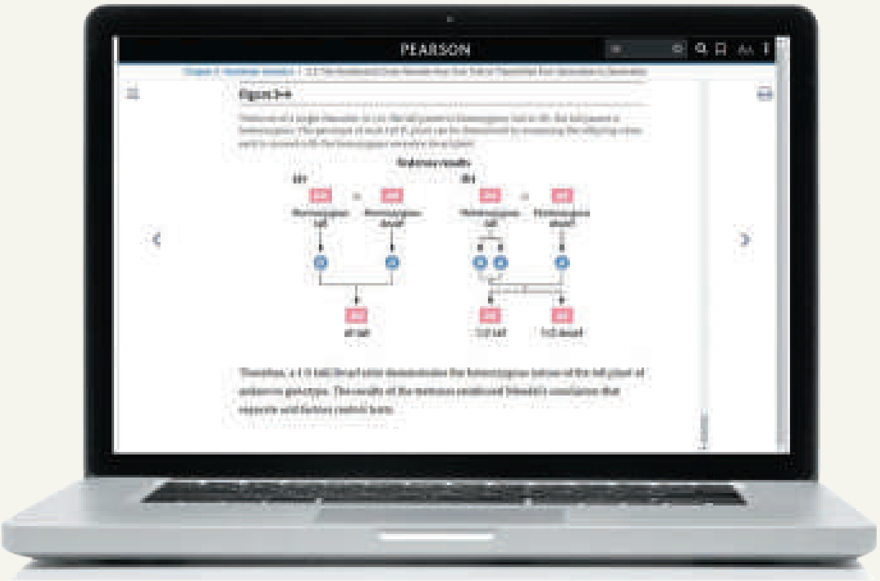
Drag the labels to the appropriate targets to identify the sequence and orientation of the daughter strand. Blue labels can be used once, more than once, or not at all.

Incorrect; Try Again

You labeled 2 of 13 targets incorrectly. U represents uracil. Note that uracil is part of a ribonucleotide and is a component of RNA, not DNA.

Access the text anytime, anywhere with Pearson eText

NEW! Pearson eText is built to adapt to any device readers are using—smartphone, tablet, or computer.



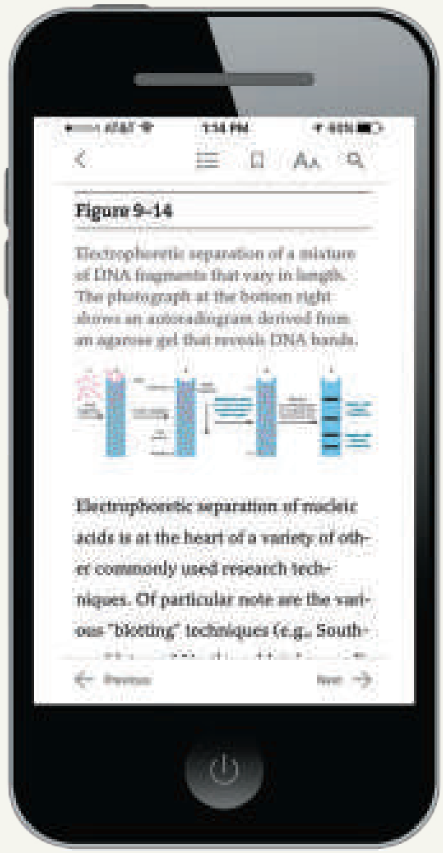


Figure 9–14

Electrophoretic separation of a mixture of DNA fragments that vary in length. The photograph at the bottom right shows an autoradiogram derived from an agarose gel that reveals DNA bands.

Electrophoretic separation of nucleic acids is at the heart of a variety of other commonly used research techniques. Of particular note are the various "blotting" techniques (e.g., South-

Pearson eText Mobile App offers offline access and can be downloaded for most iOS and Android phones/tablets from the Apple App Store or Google Play.

- Accessible (screen-reader ready)
- Configurable reading settings, including resizable type and night reading mode
- Seamlessly integrated animations
- Instructor and student note-taking, highlighting, bookmarking, and search

Contents

PART ONE
GENES, CHROMOSOMES, AND HEREDITY

7 Sex Determination and Sex Chromosomes 151

8 Chromosomal Mutations: Variation in Number and Arrangement 171

PART TWO
DNA: STRUCTURE, REPLICATION, AND ORGANIZATION

PART THREE
GENE EXPRESSION AND ITS REGULATION

PART FOUR
GENETIC TECHNOLOGY AND GENOMICS

SPECIAL TOPICS IN MODERN GENETICS 1

SPECIAL TOPICS IN MODERN GENETICS 2

SPECIAL TOPICS IN MODERN GENETICS 3

SPECIAL TOPICS IN MODERN GENETICS 4

SPECIAL TOPICS IN MODERN GENETICS 5

Gene Therapy 695

SPECIAL TOPICS IN MODERN GENETICS 6

Advances in Neurogenetics: The Study of Huntington Disease 710

Preface

It is essential that textbook authors step back and look with fresh eyes as each edition of their work is planned. In doing so, two main questions must be posed: (1) How has the body of information in their field—in this case, Genetics—grown and shifted since the last edition? (2) Which pedagogic innovations that are currently incorporated into the text should be maintained, modified, or deleted? The preparation of the 12th edition of *Concepts of Genetics*, a text well into its fourth decade of providing support for students studying in this field, has occasioned still another fresh look. And what we focused on in this new edition, in addition to the normal updating that is inevitably required, were three things:

1. **The importance of continuing to provide comprehensive coverage of important, emerging topics.**

   In this regard, we continue to include a unique approach in genetics textbooks that offers readers a set of abbreviated, highly focused chapters that we label **Special Topics in Modern Genetics**. In this edition, these provide unique, cohesive coverage of six important topics: *CRISPR-Cas and Genomic Editing*, *DNA Forensics*, *Genomics and Precision Medicine*, *Genetically Modified Foods*, *Gene Therapy*, and *Advances in Neurogenetics: The Study of Huntington Disease*. The initial and final chapters in this series are both new to this edition.

2. **The recognition of the vastly increased knowledge resulting from the study of gene regulation in eukaryotes.**

   To that end, the single chapter on this topic in previous editions has been expanded to three chapters: "Transcriptional Regulation in Eukaryotes" (Chapter 17), "Posttranscriptional Regulation in Eukaryotes" (Chapter 18), and "Epigenetic Regulation of Gene Expression" (Chapter 19). This extended coverage reflects many recent discoveries that reveal that RNA in many forms other than those that are essential to the process of transcription and translation (mRNA, tRNA, and rRNA) play critical roles in the regulation of eukaryotic gene activity. As well, it is now clear based on molecular studies related to epigenetics that this topic is best taught as an integral part of eukaryotic gene regulation. This new material provides the student exposure to modern coverage of a significant research topic.

3. **The importance of providing an increased emphasis on ethical considerations that genetics is bringing into everyday life.**

   Regarding this point, we have converted the essay feature *Genetics, Technology, and Society* to one with added emphasis on ethics and renamed it *Genetics, Ethics, and Society*. Approximately half the chapters have new or revised essays. In addition, the feature called *Case Study*, which appears near the end of all chapters, has been recast with an increased focus on ethics. Both of these features increase the opportunities for active and cooperative learning.

Goals

In the 12th edition of *Concepts of Genetics*, as in all past editions, we have five major overarching goals. Specifically, we have sought to:

- Emphasize the basic concepts of genetics.
- Write clearly and directly to students, providing understandable explanations of complex, analytical topics.
- Maintain our strong emphasis on and provide multiple approaches to problem solving.
- Propagate the rich history of genetics, which so beautifully illustrates how information is acquired during scientific investigation.
- Create inviting, engaging, and pedagogically useful full-color figures enhanced by equally helpful photographs to support concept development.

These goals collectively serve as the cornerstone of *Concepts of Genetics*. This pedagogic foundation allows the book to be used in courses with many different approaches and lecture formats.

Writing a textbook that achieves these goals and having the opportunity to continually improve on each new edition has been a labor of love for all of us. The creation of each of the twelve editions is a reflection not only of our passion for teaching genetics, but also of the constructive feedback and encouragement provided by adopters, reviewers, and our students over the past four decades.

New to This Edition

New to this edition are four chapters. Two are Special Topics in Modern Genetics entries entitled "CRISPR-Cas and Genome Editing" and "Advances in Neurogenetics: The Study of Huntington Disease." Both cover cutting-edge information and represent very recent breakthroughs in genetics. CRISPR, a genome-editing tool, is a straightforward technique that allows specific, highly accurate modification of DNA sequences within genes and is thus a powerful tool in the world of genetic research and gene therapy. In addition to this chapter, we call your attention to the introduction to Chapter 1 for an introduction to CRISPR and to also note that we have chosen this gene-editing system as the subject matter illustrated on the cover. Special Topics Chapter 6 illustrates the many of advances that have been made in the study of human neurogenetics. Huntington disease, a monogenic human disorder, has been subjected to analysis for over 40 years using every major approach and technique developed to study molecular genetics, and as such, exemplifies the growing body of information that has accrued regarding its causes, symptoms, and future treatment.

Additional new chapters arise from a major reorganization and expansion of our coverage of regulation of gene expression in eukaryotes, where we have split our previous coverage into three parts: transcriptional regulation (Chapter 17), posttranscriptional regulation (Chapter 18), and epigenetic regulation (Chapter 19). Chapter 18 includes much of the content previously contained in the Special Topics chapter *Emerging Roles of RNA* in the previous edition. Chapter 19, focused on epigenetics, is an expansion of the content previously contained in the *Epigenetics* Special Topics chapter from the previous edition.

Collectively, the addition of these four new chapters provides students and instructors with a much clearer, up-to-date presentation to these important aspects of genetics.

Continuing Pedagogic Features

We continue to include features that are distinct from, and go beyond, the text coverage, which encourage active and cooperative learning between students and the instructor.

- **Modern Approaches to Understanding Gene Function** This feature highlights how advances in genetic technology have led to our modern understanding of gene function. Appearing in many chapters, this feature prompts students to apply their analytical thinking skills, linking the experimental technology to the findings that enhance our understanding of gene function.
- **Genetics, Ethics, and Society** This feature provides a synopsis of an ethical issue related to a current finding in genetics that impacts directly on society today. It includes a section called *Your Turn*, which directs students to related resources of short readings and Web sites to support deeper investigation and discussion of the main topic of each essay.
- **Case Study** This feature, at the end of each chapter, introduces a short vignette of an everyday genetics-related situation, followed by several discussion questions. Use of the Case Study should prompt students to relate their newly acquired information in genetics to ethical issues that they may encounter away from the course.
- **Evolving Concept of the Gene** This short feature, integrated in appropriate chapters, highlights how scientists' understanding of the gene has changed over time. Since we cannot see genes, we must infer just what this unit of heredity is, based on experimental findings. By highlighting how scientists' conceptualization of the gene has advanced over time, we aim to help students appreciate the process of discovery that has led to an ever more sophisticated understanding of hereditary information.
- **How Do We Know Question** Found as the initial question in the *Problems and Discussion Questions* at the end of each chapter, this feature emphasizes the pedagogic value of studying how information is acquired in science. Students are asked to review numerous findings discussed in the chapter and to summarize the process of discovery that was involved.
- **Concept Question** This feature, found as the second question in the *Problems and Discussion Questions* at the end of each chapter, asks the student to review and comment on common aspects of the Chapter Concepts, listed at the beginning of each chapter. This feature places added emphasis on our pedagogic approach of conceptual learning.
- **Mastering Genetics** This robust online homework and assessment program guides students through complex topics in genetics, using in-depth tutorials that coach students to correct answers with hints and feedback specific to their misconceptions. New content for the 12th edition of *Concepts of Genetics* includes tutorials on emerging topics such as CRISPR-Cas, and Dyanamic Study Modules, interactive flash cards that help students master basic content so they can be more prepared for class and for solving genetics problems.

New and Updated Topics

We have revised each chapter in the text to present the most current, relevant findings in genetics. Here is a list of some of the most significant new and updated topics covered in this edition.

Chapter 1: Introduction to Genetics

- New introductory vignette that discusses the discovery and applications of the genome-editing CRISPR-Cas system
- Updated section "We Live in the Age of Genetics"

Chapter 7: Sex Determination and Sex Chromosomes

- Updated content on the XIST gene product as a long noncoding RNA
- New insights about a novel gene involved in temperature-sensitive differentiation of snapping turtles and lizards, as well as the impact of climate change on sex, sex reversal, and sex ratios

Chapter 9: Extranuclear Inheritance

- Updated information on mtDNA disorders and nuclear DNA mismatches

Chapter 11: DNA Replication and Recombination

- New coverage of the role of telomeres in disease, aging, and cancer
- New and expanded coverage of telomeres and chromosome stability, explaining how telomeres protect chromosome ends

Chapter 13: The Genetic Code and Transcription

- New coverage on transcription termination in bacteria
- New section entitled "Why Do Introns Exist?"
- Updated coverage on RNA editing

Chapter 14: Translation and Proteins

- New coverage of eukaryotic closed-loop translation, including a new figure
- Revised coverage of Beadle and Tatum's classic experiments
- Expanded coverage on the posttranslational modifications of proteins
- New coverage of the insights gleaned from the crystal structure of the human 80S ribosome

Chapter 15: Gene Mutation, DNA Repair, and Transposons

- New and revised coverage on transposons, focusing on the mechanisms of transposition by both retrotransposons and DNA transposons, as well as a discussion of how transposition creates mutations. Two new tables and five new figures are included
- Reorganization of the mutation classification section with table summaries
- New and expanded coverage of human germ-line and somatic mutation rates

Chapter 17: Transcriptional Regulation in Eukaryotes

- Revised chapter organization focuses specifically on transcriptional regulation
- Revised coverage of regulation of the *GAL* gene system in yeast with an updated figure
- New coverage on genetic boundary elements called insulators

Chapter 18: Posttranscriptional Regulation in Eukaryotes

- New chapter that greatly expands upon the previous coverage of posttranscriptional gene regulation in eukaryotes
- Revised and expanded coverage of alternative splicing and its relevance to human disease
- Expanded coverage on RNA stability and decay with a new figure
- Updated coverage of noncoding RNAs that regulate gene expression with a new figure
- Enriched coverage of ubiquitin-mediated protein degradation with a new figure

Chapter 19: Epigenetic Regulation of Gene Expression

- New chapter emphasizing the role of epigenetics in regulating gene expression, including coverage of cancer, transmission of epigenetic traits across generations, and epigenetics and behavior
- New coverage on the recently discovered phenomenon of monoallelic expression of autosomal genes
- Updated coverage of epigenome projects

Chapter 20: Recombinant DNA Technology

- Increased emphasis on the importance of whole-genome sequencing approaches
- New coverage of CRISPR-Cas as a gene editing approach, including a new figure
- Updated content on next-generation and third-generation sequencing

Chapter 21: Genomic Analysis

- Increased emphasis on the integration of genomic, bioinformatic, and proteomic approaches to analyzing genomes and understanding genome function

- A new section entitled “Genomic Analysis Before Modern Sequencing Methods,” which briefly summarizes approaches to mapping and identifying genes prior to modern sequencing
- Reorganized and revised content on the Human Genome Project. Updated content on personal genome projects and new content on diploid genomes and mosaicism and the pangenome to emphasize human genetic variations
- New coverage of the Human Microbiome Project including a new figure displaying microbiome results of patients with different human disease conditions
- New coverage of *in situ* RNA sequencing

Chapter 22: Applications of Genetic Engineering and Biotechnology

- Updated content on biopharmaceutical products including newly approved recombinant proteins, DNA vaccine trials to immunize against Zika virus, genetically modified organisms, and gene drive in mosquitos to control the spread of Zika
- New coverage of genes essential for life and how synthetic genomics is being applied to elucidate them. Clarification of prognostic and diagnostic genetics tests and the relative value of each for genetic analysis
- New content on DNA and RNA sequencing
- New section entitled “Screening the Genome for Genes or Mutations You Want,” which discusses how scientists can look at genetic variation that confers beneficial phenotypes
- New section entitled “Genetic Analysis by Personal Genomics Can Include Sequencing of DNA and RNA” that expands coverage of personal genome projects and new approaches for single-cell genetic analysis of DNA and RNA

Chapter 23: Developmental Genetics

- New section entitled “Epigenetic Regulation of Development”
- New coverage of DNA methylation and progressive restriction of developmental potential
- Expanded coverage of binary switch genes and regulatory networks

Chapter 24: Cancer Genetics

- Extended coverage of environmental agents that contribute to human cancers, including more information about both natural and human-made carcinogens
- New section entitled “Tobacco Smoke and Cancer” explaining how a well-studied carcinogen induces a wide range of genetic effects that may lead to mutations and cancer
- New section entitled “Cancer Therapies and Cancer Cell Biology,” describing the mechanisms of chemotherapies and radiotherapies as they relate to cancer cell proliferation, DNA repair, and apoptosis

Chapter 25: Quantitative Genetics and Multifactorial Traits

- Updated coverage on quantitative trait loci (QTLs)
- Revised and expanded section entitled “eQTLs and Gene Expression”

Chapter 26: Population and Evolutionary Genetics

- New coverage on vertebrate evolution
- New coverage of phylogenetic trees
- Updated coverage on the origins of the human genome
- New section entitled “Genotype and Allele Frequency Changes”
- New coverage on pre- and post-zygotic isolating mechanisms

Special Topic Chapter 1: CRISPR-Cas and Genome Editing

- New chapter on a powerful genome editing tool called CRISPR-Cas
- Up-to-date coverage on CRISPR-Cas applications, the patenting of this technology, and the ethical concerns of human genome editing

Special Topic Chapter 2: DNA Forensics

- New section on the still controversial DNA phenotyping method, including new explanations of how law-enforcement agencies currently use this technology

Special Topic Chapter 3: Genomics and Precision Medicine

- New section entitled “Precision Oncology,” including descriptions of two targeted cancer immunotherapies: adoptive cell transfer and engineered T-cell therapies
- Updated pharmacogenomics coverage, including a description of new trends in preemptive gene screening for pharmacogenomic variants as well as the pGEN4Kids program, a preemptive gene screening program that integrates DNA analysis data into patient electronic health records

Special Topic Chapter 4: Genetically Modified (GM) Foods

- New section entitled “Gene Editing and GM Foods” describing how scientists are using the new techniques of gene editing (including ZFN, TALENS, and CRISPR-Cas) to create GM food plants and animals, and how these methods are changing the way in which GM foods are being regulated

- A new box entitled "The New CRISPR Mushroom" describing the development and regulatory approval of the first CRISPR-created GM food to be approved for human consumption

Special Topic Chapter 5: Gene Therapy

- Updated coverage of gene therapy trials currently underway
- Reordered chapter content to highlight emergence of CRISPR-Cas in a new section entitled "Gene Editing"
- Substantially expanded content on CRISPR-Cas including a brief summary of some of the most promising trials in humans and animals to date
- Incorporation of antisense RNA and RNA interference into a new section entitled "RNA-based Therapeutics," including updated trials involving spinal muscular atrophy
- Updated content on roles for stem cells in gene therapy
- New content on combining gene editing with immunotherapy
- New ethical discussions on CRISPR-Cas and germline and embryo editing

Special Topic Chapter 6: Advances in Neurogenetics: The Study of Huntington Disease (HD)

- New chapter that surveys the study of HD commencing around 1970 up to the current time
- Coverage of the genetic basis and expression of HD, the mapping and isolation of the gene responsible for the disorder, the mutant gene product, molecular and cellular alterations caused by the mutation, transgenic animal models of HD, cellular and molecular approaches to therapy, and a comparison of HD to other inherited neurodegenerative disorders

Strengths of This Edition

- **Organization** —We have continued to attend to the organization of material by arranging chapters within major sections to reflect changing trends in genetics. Of particular note is the expansion of our coverage of the regulation of gene expression in eukaryotes, now reorganized into three chapters at the end of Part Three. Additionally, Part Four continues to provide organized coverage of genomics into three carefully integrated chapters.

- **Active Learning** —A continuing goal of this book is to provide features within each chapter that small groups of students can use either in the classroom or as assignments outside of class. Pedagogic research continues to support the value and effectiveness of such active and cooperative learning experiences. To this end, there are four features that greatly strengthen this edition: *Case Study*; *Genetics, Ethics, and Society*; *Exploring Genomics*; and *Modern Approaches to Understanding Gene Function*. Whether instructors use these activities as active learning in the classroom or as assigned interactions outside of the classroom, the above features will stimulate the use of current pedagogic approaches during student' learning. The activities help engage students, and the content of each feature ensures that they will become knowledgeable about cutting-edge topics in genetics.

Emphasis on Concepts

The title of our textbook—*Concepts of Genetics*—was purposefully chosen, reflecting our fundamental pedagogic approach to teaching and writing about genetics. However, the word "concept" is not as easy to define as one might think. Most simply put, we consider a concept to be *a cognitive unit of meaning—an abstract representation that encompasses a related set of scientifically derived findings and ideas*. Thus, a concept provides a broad mental image that, for example, might reflect a straightforward snapshot in your mind's eye of what constitutes a chromosome; a dynamic vision of the detailed processes of replication, transcription, and translation of genetic information; or just an abstract perception of varying modes of inheritance.

We think that creating such mental imagery is the very best way to teach science, in this case, genetics. Details that might be memorized, but soon forgotten, are instead subsumed within a conceptual framework that is easily retained and nearly impossible to forget. Such a framework may be expanded in content as new information is acquired and may interface with other concepts, providing a useful mechanism to integrate and better understand related processes and ideas. An extensive set of concepts may be devised and conveyed to eventually encompass and represent an entire discipline—and this is our goal in this genetics textbook.

To aid students in identifying the conceptual aspects of a major topic, each chapter begins with a section called *Chapter Concepts*, which identifies the most important topics about to be presented. Each chapter ends with a section called *Summary Points*, which enumerates the five to ten key points that have been discussed. And in the *How Do We Know*? question that starts each chapter's problem set, students are asked to connect concepts to experimental findings. This question is then followed by a *Concept Question*, which asks the student to review and comment on common aspects of the Chapter Concepts. Collectively, these features help to ensure that students engage in, become aware of, and understand the major conceptual issues as they confront the extensive vocabulary and the many important details of genetics. Carefully designed figures also support our conceptual approach throughout the book.

Emphasis on Problem Solving

As authors and teachers, we have always recognized the importance of enhancing students' problem-solving skills. Students need guidance and practice if they are to develop into strong analytical thinkers. To that end, we present a suite of features in every chapter to optimize opportunities for student growth in the important areas of problem solving and analytical thinking.

- **Now Solve This** Found several times within the text of each chapter, each entry provides a problem similar to ones found at the end of the chapter that is closely related to the current text discussion. In each case, a pedagogic hint is provided to offer insight and to aid in solving the problem.

- **Insights and Solutions** As an aid to the student in learning to solve problems, the *Problems and Discussion Questions* section of each chapter is preceded by what has become an extremely popular and successful section. *Insights and Solutions* poses problems or questions and provides detailed solutions and analytical insights as answers are provided. The questions and their solutions are designed to stress problem solving, quantitative analysis, analytical thinking, and experimental rationale. Collectively, these constitute the cornerstone of scientific inquiry and discovery.

- **Problems and Discussion Questions** Each chapter ends with an extensive collection of *Problems and Discussion Questions*. These include several levels of difficulty, with the most challenging (*Extra-Spicy Problems*) located at the end of each section. Often, Extra-Spicy Problems are derived from the literature of genetic research, with citations. Brief answers to all even-numbered problems are presented in Appendix B. The *Student Handbook and Solutions Manual* answers every problem and is available to students whenever faculty decide that it is appropriate.

- **How Do We Know?** Appearing as the first entry in the *Problems and Discussion Questions* section, this question asks the student to identify and examine the experimental basis underlying important concepts and conclusions that have been presented in the chapter. Addressing these questions will aid the student in more fully understanding, rather than memorizing, the endpoint of each body of research. This feature is an extension of the learning approach in biology first formally described by John A. Moore in his 1999 book *Science as a Way of Knowing—The Foundation of Modern Biology*.

- **Mastering Genetics** Tutorials in Mastering Genetics help students strengthen their problem-solving skills while exploring challenging activities about key genetics content. In addition, end-of-chapter problems are also available for instructors to assign as online homework. Students will also be able to access materials in the Study Area that help them assess their understanding and prepare for exams.

For the Instructor

Mastering Genetics—http://www.masteringgenetics.com

Mastering Genetics engages and motivates students to learn and allows you to easily assign automatically graded activities. Tutorials provide students with personalized coaching and feedback. Using the gradebook, you can quickly monitor and display student results. Mastering Genetics easily captures data to demonstrate assessment outcomes. Resources include:

- New Dynamic Study Modules, which are interactive flashcards, provide students with multiple sets of questions with extensive feedback so they can test, learn, and retest until they achieve mastery of the textbook material. These can be assigned for credit or used for self-study, and they are powerful preclass activities that help prepare students for more involved content coverage or problem solving in class.

- New tutorials on topics like CRISPR-Cas will help students master important, challenging concepts.

- In-depth tutorials that coach students with hints and feedback specific to their misconceptions

- An item library of thousands of assignable questions including end-of-chapter problems, reading quizzes, and test bank items. You can use publisher-created prebuilt assignments to get started quickly. Each question can be easily edited to match the precise language you use.

- Over 100 Practice Problems are like end-of-chapter questions in scope and level of difficulty and are found only in Mastering Genetics. Solutions are not available in the Student Solutions Manual, and the bank of questions extends your options for assigning challenging problems. Each problem includes specific wrong answer feedback to help students learn from their mistakes and to guide them toward the correct answer.

- eText 2.0 provides a dynamic digital version of the textbook, including embedded videos. The text adapts to the size of the screen being used, and features include student and instructor note-taking, highlighting, bookmarking, search, and hot-linked glossary.

- A gradebook that provides you with quick results and easy-to-interpret insights into student performance.

Downloadable Instructor Resources

The Instructor Resources for the 12th edition offers adopters of the text convenient access to a comprehensive and innovative set of lecture presentation and teaching tools. Developed to meet the needs of veteran and newer instructors alike, these resources include:

- The JPEG files of all text line drawings with labels individually enhanced for optimal projection results (as well as unlabeled versions) and all text tables.
- Most of the text photos, including all photos with pedagogical significance, as JPEG files.
- The JPEG files of line drawings, photos, and tables preloaded into comprehensive PowerPoint® presentations for each chapter.
- A second set of PowerPoint® presentations consisting of a thorough lecture outline for each chapter augmented by key text illustrations.
- PowerPoint® presentations containing a comprehensive set of in-class clicker questions for each chapter.
- An impressive series of concise instructor animations adding depth and visual clarity to the most important topics and dynamic processes described in the text.
- In Word and PDF files, a complete set of the assessment materials and study questions and answers from the testbank. Files are also available in TestGen format.

TestGen EQ Computerized Testing Software

(013483223X/9780134832234) Test questions are available as part of the TestGen EQ Testing Software, a text-specific testing program that is networkable for administering tests. It also allows instructors to view and edit questions, export the questions as tests, and print them out in a variety of formats.

For the Student

Student Handbook and Solutions Manual

(0134870085 / 9780134870083) Authored by Michelle Gaudette (*Tufts University*) and Harry Nickla (*Creighton University*-Emeritus). This valuable handbook provides a detailed step-by-step solution or lengthy discussion for every problem in the text. The handbook also features additional study aids, including extra study problems, chapter outlines, vocabulary exercises, and an overview of how to study genetics.

Mastering Genetics—http://www.masteringgenetics.com

Used by over one million science students, the Mastering platform is the most effective and widely used online tutorial, homework, and assessment system for the sciences; it helps students perform better on homework and exams. As an instructor-assigned homework system, Mastering Genetics is designed to provide students with a variety of assessment tools to help them understand key topics and concepts and to build problem-solving skills. Mastering Genetics tutorials guide students through the toughest topics in genetics with self-paced tutorials that provide individualized coaching with hints and feedback specific to a student's individual misconceptions. Students can also explore the Mastering Genetics Study Area, which includes animations, the eText, *Exploring Genomics* exercises, and other study aids. The interactive eText 2.0 allows students to highlight text, add study notes, review instructor's notes, and search throughout the text.

Acknowledgments

Contributors

We begin with special acknowledgments to those who have made direct contributions to this text. First of all, we are pleased to acknowledge the work of Michelle Gaudette, who has assumed responsibility for writing the *Student Handbook and Solutions Manual* and the answers in Appendix B. We much appreciate this important contribution. We also thank Jutta Heller of the University of Washington—Tacoma, Christopher Halweg of North Carolina State University, Pamela Osenkowski of Loyola University—Chicago, and John Osterman of the University of Nebraska—Lincoln for their work on the media program. Steven Gorsich of Central Michigan University, Virginia McDonough of Hope College, Cindy Malone of California State University—Northridge, Pamela Marshall of Arizona State University West, and Brad Mehrtens of University of Illinois all made important contributions to the instructor resources program. We are grateful to all of these contributors not only for sharing their genetic expertise, but for their dedication to this project as well as the pleasant interactions they provided.

Proofreaders and Accuracy Checking

Reading the manuscript of an 800+ page textbook deserves more thanks than words can offer. Our utmost appreciation is extended to Ann Blakey, *Ball State University*, Jutta Heller, *University of Washington—Tacoma*, and Valerie Oke, *University of Pittsburgh*, who provided accuracy checking of many chapters, and to Kay Brimeyer, who proofread the entire manuscript. They confronted this task with patience and diligence, contributing greatly to the quality of this text.

Reviewers

All comprehensive texts are dependent on the valuable input provided by many reviewers. While we take full responsibility for any errors in this book, we gratefully acknowledge the help provided by those individuals who reviewed the content and pedagogy of this edition:

Kyle Belozerov, York University
Ann Blakey, Ball State University
Michelle Boissiere, Xavier University of Louisiana
Gregory Booton, The Ohio State University
Mary Bryk, Texas A&M University
Soochin Cho, Creighton University
Claire Cronmiller, University of Virginia
Johnny El-Rady, University of South Florida
Christy Fillman, University of Colorado—Boulder
Edison Fowlks, Hampton University
Shaghayegh Harbi, New York University
Christopher Harendza, Montgomery County Community College
Benjamin Harrison, University of Alaska—Anchorage
Steven Karpowicz, University of Central Oklahoma
David Kass, Eastern Michigan University
Oliver Kerscher, The College of William and Mary
Kirkwood Land, University of the Pacific
Howard Laten, Loyola University Chicago
Haiying Liang, Clemson University
Te-Wen Lo, Ithaca College
Kari Loomis, University of Massachusetts—Amherst
TyAnna Lovato, University of New Mexico
Matthew Marcello, Pace University
Pamela Marshall, Arizona State University West
Jeff Maughan, Brigham Young University
Virginia McDonough, Hope College
Craig Miller, University of California—Berkeley
Anna Newman, University of Houston
Pamela Osenkowski, Loyola University Chicago
Stephen Page, Community College of Baltimore County
Dennis Ray, The University of Arizona
Jennifer Schisa, Central Michigan University
Randall Small, The University of Tennessee—Knoxville
Paul Small, Eureka College
Doug Thrower, University of California—Santa Barbara
Harald Vaessin, The Ohio State University
Meena Vijayaraghavan, Tulane University

As these acknowledgments make clear, a text such as this is a collective enterprise. All of the individuals above deserve to share in the success this text enjoys. We want them to know that our gratitude is equaled only by the extreme dedication evident in their efforts. Many, many thanks to them all.

Editorial and Production Input

At Pearson, we express appreciation and high praise for the editorial guidance of Michael Gillespie, whose ideas and efforts have helped to shape and refine the features of this edition of the text. Brett Coker, our Content Producer, has worked tirelessly to keep the project on schedule and to maintain our standards of high quality. Sonia DiVittorio provided detailed feedback in her role as developmental editor. In addition, our editorial team—Ginnie Simione Jutson, Director of Content Development; Barbara Price, Senior Development Editor; Margot Otway, Senior Development Editor; Sarah Jensen, Senior Content Developer for Mastering Genetics; and Chloe Veylit, Rich Media Producer—have provided valuable input into the current edition. They have worked creatively to ensure that the pedagogy and design of the book and media package are at the cutting edge of a rapidly changing discipline. Summer Giles, Editorial Assistant, helped manage and edit the instructor resources. Shaghayegh Harbi authored engaging, insightful tutorials for Mastering Genetics. Michelle Gardner supervised all the production intricacies with great attention to detail and perseverance. Outstanding copyediting was performed by Lucy Mullins, for which we are most grateful. Kelly Galli and Christa Pelaez have professionally and enthusiastically managed the marketing of the text. Finally, the beauty and consistent presentation of the artwork are the product of Imagineering of Toronto. Without the work ethic and dedication of the above individuals, the text would never have come to fruition.

1

Introduction to Genetics

Newer model organisms in genetics include the roundworm, *Caenorhabditis elegans*; the zebrafish, *Danio rerio*; and the mustard plant, *Arabidopsis thaliana*.

CHAPTER CONCEPTS

- Genetics in the twenty-first century is built on a rich tradition of discovery and experimentation stretching from the ancient world through the nineteenth century to the present day.
- Transmission genetics is the general process by which traits controlled by genes are transmitted through gametes from generation to generation.
- Mutant strains can be used in genetic crosses to map the location and distance between genes on chromosomes.
- The Watson–Crick model of DNA structure explains how genetic information is stored and expressed. This discovery is the foundation of molecular genetics.
- Recombinant DNA technology revolutionized genetics, was the foundation for the Human Genome Project, and has generated new fields that combine genetics with information technology.
- Biotechnology provides genetically modified organisms and their products that are used across a wide range of fields including agriculture, medicine, and industry.
- Model organisms used in genetics research are now utilized in combination with recombinant DNA technology and genomics to study human diseases.
- Genetic technology is developing faster than the policies, laws, and conventions that govern its use.

One of the small pleasures of writing a genetics textbook is being able to occasionally introduce in the very first paragraph of the initial chapter a truly significant breakthrough in the discipline that hopefully will soon have a major, diverse impact on human lives. In this edition, we are fortunate to be able to discuss the discovery of **CRISPR-Cas**, a molecular complex found in bacteria that has the potential to revolutionize our ability to rewrite the DNA sequence of genes from any organism. As such, it represents the ultimate tool in genetic technology, whereby the genome of organisms, including humans, may be precisely edited. Such gene modification represents the ultimate application of the many advances in biotechnology made in the last 35 years, including the sequencing of the human genome.

Other systems have been developed, including **zinc-finger nucleases (ZFNs)** and **transcription activator-like effector nucleases (TALENs)**, that are now undergoing clinical trials for the treatment of human diseases, and which we will discuss later in the text. However, the CRISPR-Cas system is the most powerful and far-reaching method and is now the preferred approach in gene modification. This system allows researchers to edit genomes with greater accuracy, is easier to use, and is more versatile than the ZFN or TALEN systems. CRISPR-Cas molecules were initially discovered as a molecular complex that protects bacterial cells from invasion by viruses. CRISPR (clustered regularly interspersed short palindromic repeats) designates an RNA molecule, which in the laboratory can be synthesized to match any DNA sequence of choice. CRISPR RNA has two ends: one recognizes and binds to a matching DNA sequence in the gene of interest, and the other binds to a CRISPR-associated (Cas) nuclease, or DNA-cutting enzyme. The most commonly used Cas nuclease is Cas9, but there are many other Cas nucleases, each of which has slightly different properties, contributing to the system's versatility. In laboratory experiments, CRISPR-Cas systems have already been used to repair mutations in cells derived from individuals with several genetic disorders, including cystic fibrosis, Huntington disease, beta-thalassemia, sickle cell disease, muscular dystrophy, and X-linked retinitis pigmentosa, which results in progressive vision loss. In the United States a clinical trial using CRISPR-Cas9 for genome editing in cancer therapy has been approved, and a second proposal for treating a genetic form of blindness is in preparation. A clinical trial using CRISPR-Cas9 for cancer therapy is already under way in China.

The application of this remarkable system goes far beyond research involving human genetic disorders. In organisms of all kinds, wherever genetic modification may improve on nature to the benefit of human existence and of our planet, the use of CRISPR-Cas will find many targets. For example, one research group was able to use this system to spread genes that prevent mosquitoes from carrying the parasite that causes malaria. Other researchers have proposed using CRISPR-Cas9 to engineer laboratory-grown human blood vessels and organs that do not express proteins that cause rejection of transplanted tissues and organs. The method has also been used to create disease-resistant strains of wheat and rice.

The power of this system, like any major technological advance, has already raised ethical concerns. For example, genetic modification of human germ cells or embryos would change the genetic information carried by future generations. These modifications may have unintended and significant negative consequences for our species. An international summit on human gene editing in December 2015 concluded that a global forum to address concerns about heritable modifications should be convened to formulate regulations that apply to all countries involved in CRISPR research.

CRISPR-Cas may turn out to be one of the most exciting genetic advances in decades. We will return later in the text to an extended discussion of its discovery, describe how it works, its many applications, and the ethical considerations that it raises (see Special Topic Chapter 1—CRISPR and Genomic Editing).

For now, we hope that this short introduction has stimulated your curiosity, interest, and enthusiasm for the study of genetics. The remainder of this chapter provides an overview of major concepts of genetics and a survey of the major turning points in the history of the discipline. Along the way, enjoy your studies, but take your responsibilities as a novice geneticist most seriously.

1.1 Genetics Has a Rich and Interesting History

We don't know when people first recognized the hereditary nature of certain traits, but archaeological evidence (e.g., pictorial representations, preserved bones and skulls, and dried seeds) documents the successful domestication of animals and the cultivation of plants thousands of years ago by the artificial selection of genetic variants from wild populations. Between 8000 and 1000 B.C., horses, camels, oxen, and wolves were domesticated, and selective breeding of these species soon followed. Cultivation of many plants, including maize, wheat, rice, and the date palm, began around 5000 B.C. Such evidence documents our ancestors' successful attempts to manipulate the genetic composition of species.

During the Golden Age of Greek culture, the writings of the Hippocratic School of Medicine (500–400 B.C.) and of the philosopher and naturalist Aristotle (384–322 B.C.) discussed heredity as it relates to humans. The Hippocratic treatise *On the Seed* argued that active "humors" in various parts of the body served as the bearers of hereditary traits. Drawn from various parts of the male body to the semen and passed on to offspring, these humors could be healthy or diseased, with the diseased humors accounting for the appearance of newborns with congenital disorders or deformities. It was also believed that these humors could be altered in individuals before they were passed on to offspring, explaining how newborns could "inherit" traits that their parents had "acquired" in response to their environment.

Aristotle extended Hippocrates' thinking and proposed that the male semen contained a "vital heat" with the capacity to produce offspring of the same "form" (i.e., basic structure and capacities) as the parent. Aristotle believed that this heat cooked and shaped the menstrual blood produced by the female, which was the "physical substance" that gave rise to an offspring. The embryo developed not because it already contained the parts of an adult in miniature form (as some Hippocratics had thought) but because of the shaping power of the vital heat. Although the ideas of Hippocrates and Aristotle sound primitive and naive today, we should recall that prior to the 1800s neither sperm nor eggs had been observed in mammals.

1600–1850: The Dawn of Modern Biology

Between about 300 B.C. and 1600 A.D., there were few significant new ideas about genetics. However, between 1600 and 1850, major strides provided insight into the biological basis of life. In the 1600s, William Harvey studied reproduction and development and proposed the theory of **epigenesis**, which states that an organism develops from the fertilized egg by a succession of developmental events that eventually transform the egg into an adult. The theory of epigenesis directly conflicted with the theory of **preformation**, which stated that the fertilized egg contains a complete miniature adult, called a **homunculus** (**Figure 1.1**). Around 1830, Matthias Schleiden and Theodor Schwann proposed the **cell theory**, stating that all organisms are composed of basic structural units called cells, which are derived from preexisting cells. The idea of **spontaneous generation**, the creation of living organisms from nonliving components, was disproved by Louis Pasteur later in the century, and living organisms were then considered to be derived from preexisting organisms and to consist of cells.

In the mid-1800s the revolutionary work of Charles Darwin and Gregor Mendel set the stage for the rapid development of genetics in the twentieth and twenty-first centuries.

Charles Darwin and Evolution

With this background, we turn to a brief discussion of the work of Charles Darwin, who published *The Origin of Species*, in 1859, describing his ideas about evolution. Darwin's geological, geographical, and biological observations convinced him that existing species arose by descent with modification from ancestral species. Greatly influenced by his voyage on the HMS *Beagle* (1831–1836), Darwin's thinking led him to formulate the theory of **natural selection**, which presented an explanation of the mechanism of evolutionary change. Formulated and proposed independently by Alfred Russel Wallace, natural selection is based on the observation that populations tend to contain more offspring than the environment can support, leading to a struggle for survival among individuals. Those individuals with heritable traits that allow them to adapt to their environment are better able to survive and reproduce than those with less adaptive traits. Over a long period of time, advantageous variations, even very slight ones, will accumulate. If a population carrying these inherited variations becomes reproductively isolated, a new species may result.

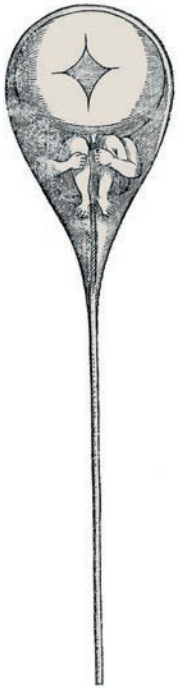


FIGURE 1.1 Depiction of the *homunculus*, a sperm containing a miniature adult, perfect in proportion and fully formed.

Darwin, however, lacked an understanding of the genetic basis of variation and inheritance, a gap that left his theory open to reasonable criticism well into the twentieth century. Shortly after Darwin published his book, Gregor Johann Mendel published a paper in 1866 showing how traits were passed from generation to generation in pea plants and offering a general model of how traits are inherited. His research was little known until it was partially duplicated and brought to light by Carl Correns, Hugo de Vries, and Erich Tschermak around 1900.

By the early part of the twentieth century, it became clear that heredity and development were dependent on genetic information residing in genes contained in chromosomes, which were then contributed to each individual by gametes—the so-called *chromosomal theory of inheritance*. The gap in Darwin's theory was closed, and Mendel's research has continued to serve as the foundation of genetics.

1.2 Genetics Progressed from Mendel to DNA in Less Than a Century

Because genetic processes are fundamental to life itself, the science of genetics unifies biology and serves as its core. The starting point for this branch of science was a monastery garden in central Europe in the late 1850s.

Mendel's Work on Transmission of Traits

Gregor Mendel, an Augustinian monk, conducted a decade-long series of experiments using pea plants. He applied quantitative data analysis to his results and showed that traits are passed from parents to offspring in predictable ways.

He further concluded that each trait in the plant is controlled by a pair of factors (which we now call genes) and that during gamete formation (the formation of egg cells and sperm), members of a gene pair separate from each other. His work was published in 1866 but was largely unknown until it was cited in papers published by others around 1900. Once confirmed, Mendel's findings became recognized as explaining the transmission of traits in pea plants and all other higher organisms. His work forms the foundation for **genetics**, which is defined as the branch of biology concerned with the study of heredity and variation. Mendelian genetics will be discussed later in the text (see Chapters 3 and 4).

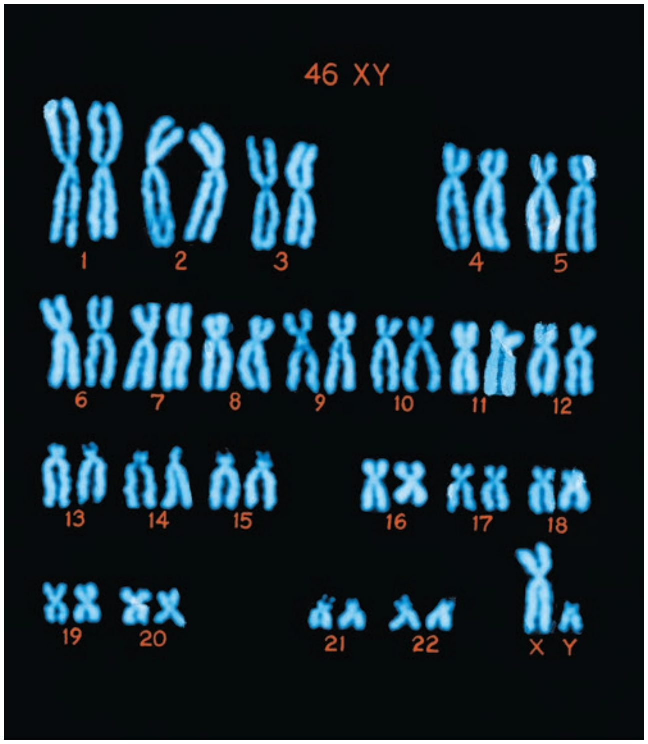


FIGURE 1.3 A colorized image of the human male chromosome set. Arranged in this way, the set is called a karyotype.

The Chromosome Theory of Inheritance: Uniting Mendel and Meiosis

Mendel did his experiments before the structure and role of chromosomes were known. About 20 years after his work was published, advances in microscopy allowed researchers to identify chromosomes (Figure 1.2) and establish that, in most eukaryotes, members of each species have a characteristic number of chromosomes called the **diploid number (2*n*)** in most of their cells. For example, humans have a diploid number of 46 (Figure 1.3). Chromosomes in diploid cells exist in pairs, called **homologous chromosomes**.

Researchers in the last decades of the nineteenth century also described chromosome behavior during two forms of cell division, **mitosis** and **meiosis**. In mitosis (Figure 1.4), chromosomes are copied and distributed so that each daughter cell receives a diploid set of chromosomes identical to those in the parental cell. Meiosis is associated with gamete formation. Cells produced by meiosis receive only one chromosome from each chromosome pair, and the resulting number of chromosomes is called the **haploid number (*n*)**. This reduction in chromosome number is essential if the offspring arising from the fusion of egg and sperm are to maintain the constant number of chromosomes characteristic of their parents and other members of their species.

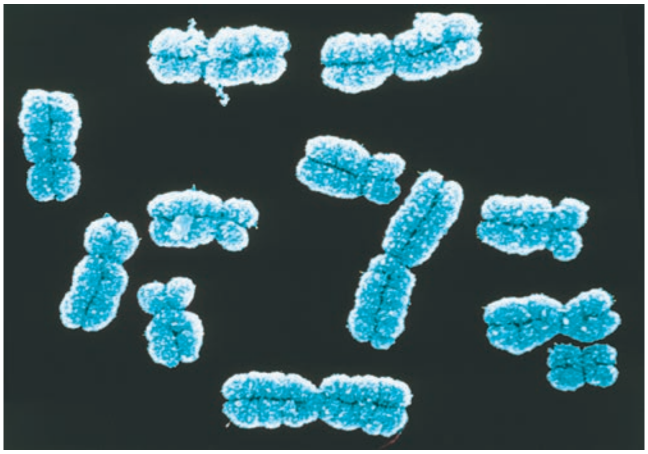


FIGURE 1.2 A colorized image of human chromosomes that have duplicated in preparation for cell division, as visualized using a scanning electron microscope.

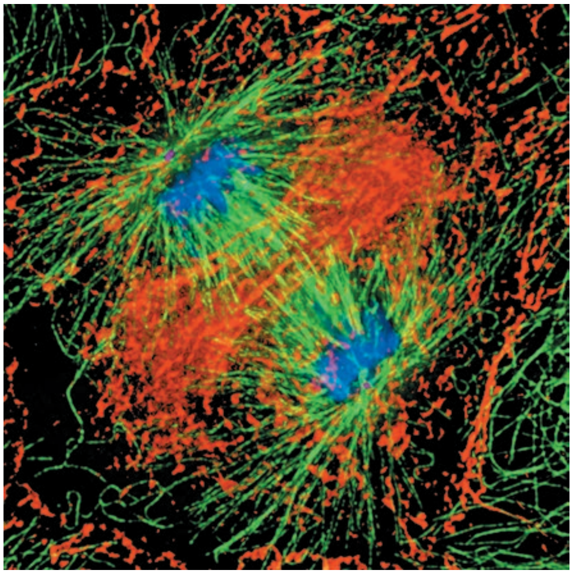


FIGURE 1.4 A late stage in mitosis after the chromosomes (stained blue) have separated.

Early in the twentieth century, Walter Sutton and Theodor Boveri independently noted that the behavior

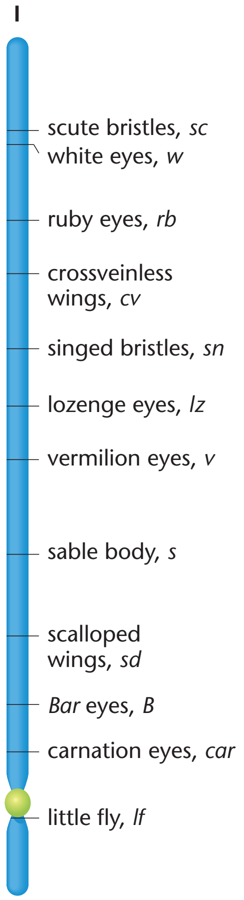


FIGURE 1.5 A drawing of chromosome I (the X chromosome, one of the sex-determining chromosomes) of *D. melanogaster*, showing the location of several genes. Chromosomes can contain hundreds of genes.

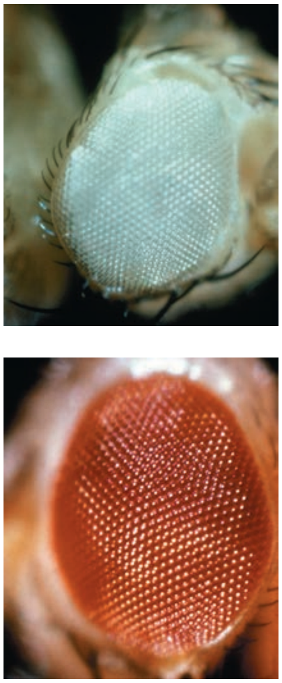


FIGURE 1.6 The white-eyed mutation in *D. melanogaster* (top) and the normal red eye color (bottom).

of chromosomes during meiosis is identical to the behavior of genes during gamete formation described by Mendel. For example, genes and chromosomes exist in pairs, and members of a gene pair and members of a chromosome pair separate from each other during gamete formation. Based on these and other parallels, Sutton and Boveri each proposed that genes are carried on chromosomes (**Figure 1.5**). They independently formulated the **chromosome theory of inheritance**, which states that inherited traits are controlled by genes residing on chromosomes faithfully transmitted through gametes, maintaining genetic continuity from generation to generation.

Genetic Variation

About the same time that the chromosome theory of inheritance was proposed, scientists began studying the inheritance of traits in the fruit fly, *Drosophila melanogaster*. Early in this work, a white-eyed fly (**Figure 1.6**) was discovered among normal (wild-type) red-eyed flies. This variation was produced by a **mutation** in one of the genes controlling eye color. Mutations are defined as any heritable change in the DNA sequence and are the source of all genetic variation.

The white-eye variant discovered in *Drosophila* is an **allele** of a gene controlling eye color. Alleles are defined as alternative forms of a gene. Different alleles may produce differences in the observable features, or **phenotype**, of an organism. The set of alleles for a given trait carried by an organism is called the **genotype**. Using mutant genes as markers, geneticists can map the location of genes on chromosomes (Figure 1.5).

The Search for the Chemical Nature of Genes: DNA or Protein?

Work on white-eyed *Drosophila* showed that the mutant trait could be traced to a single chromosome, confirming the idea that genes are carried on chromosomes. Once this relationship was established, investigators turned their attention to identifying which chemical component of chromosomes carries genetic information. By the 1920s, scientists knew that proteins and DNA were the major chemical components of chromosomes. There are a large number of different proteins, and because of their universal distribution in the nucleus and cytoplasm, many researchers thought proteins were the carriers of genetic information.

In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty, researchers at the Rockefeller Institute in New York, published experiments showing that DNA was the carrier of genetic information in bacteria. This evidence, though clear-cut, failed to convince many influential scientists. Additional evidence for the role of DNA as a carrier of genetic information came from Hershey and Chase who worked with viruses. This evidence that DNA carries genetic information, along with other research over the next few years, provided solid proof that DNA, not protein, is the genetic material, setting the stage for work to establish the structure of DNA.

1.3 Discovery of the Double Helix Launched the Era of Molecular Genetics

Once it was accepted that DNA carries genetic information, efforts were focused on deciphering the structure of the DNA molecule and the mechanism by which information stored in it produces a phenotype.

The Structure of DNA and RNA

One of the great discoveries of the twentieth century was made in 1953 by James Watson and Francis Crick, who described the structure of DNA. DNA is a long, ladder-like macromolecule that twists to form a double helix (**Figure 1.7**). Each linear strand of the helix is made up of subunits called **nucleotides**. In DNA, there are four different nucleotides, each of which contains a nitrogenous base, abbreviated A (adenine), G (guanine), T (thymine), or C (cytosine). These four bases, in various sequence combinations, ultimately encode genetic information. The two strands of DNA are exact complements of one another, so that the rungs of the ladder in the double helix always consist of A=T and G≡C base pairs. Along with Maurice Wilkins, Watson and Crick were awarded a Nobel Prize in 1962 for their work on the structure of DNA. We will discuss the structure of DNA later in the text (see Chapter 9).

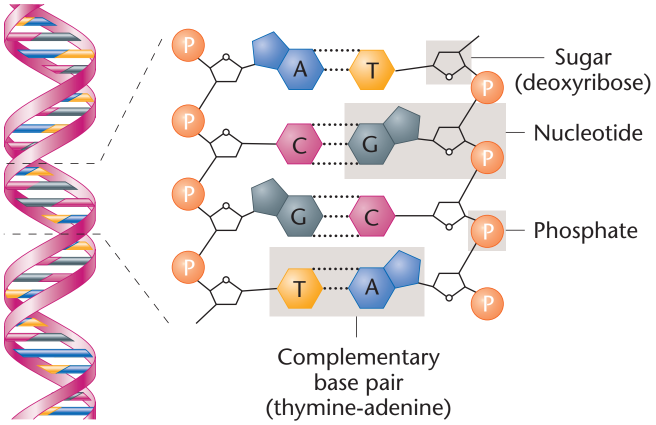


FIGURE 1.7 Summary of the structure of DNA, illustrating the arrangement of the double helix (on the left) and the chemical components making up each strand (on the right). The dotted lines on the right represent weak chemical bonds, called hydrogen bonds, which hold together the two strands of the DNA helix.

Another nucleic acid, RNA, is chemically similar to DNA but contains a different sugar (ribose rather than deoxyribose) in its nucleotides and contains the nitrogenous base uracil in place of thymine. RNA, however, is generally a single-stranded molecule.

Gene Expression: From DNA to Phenotype

The genetic information encoded in the order of nucleotides in DNA is expressed in a series of steps that results in the formation of a functional gene product. In the majority of cases, this product is a protein. In eukaryotic cells, the process leading to protein production begins in the nucleus with **transcription**, in which the nucleotide sequence in one strand of DNA is used to construct a complementary RNA sequence (top part of **Figure 1.8**). Once an RNA molecule is produced, it moves to the cytoplasm, where the RNA—called **messenger RNA**, or **mRNA** for short—binds to a **ribosome**. The synthesis of proteins under the direction of mRNA is called **translation** (center part of Figure 1.8). The information encoded in mRNA (called the **genetic code**) consists of a linear series of nucleotide triplets. Each triplet, called a **codon**, is complementary to the information stored

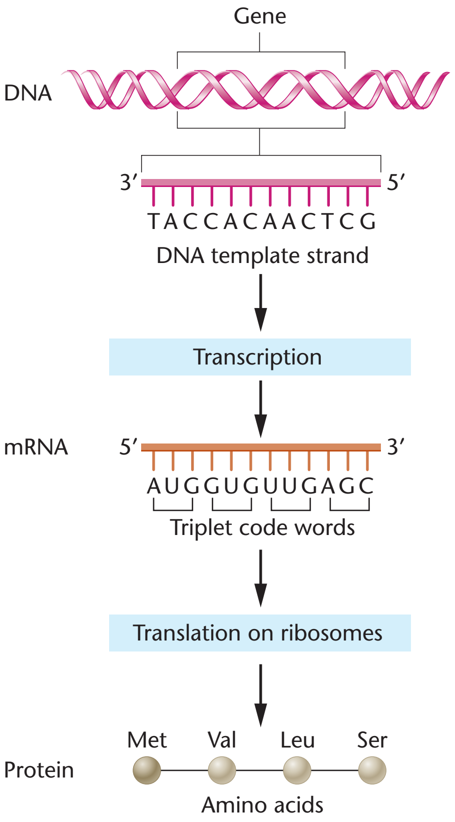


FIGURE 1.8 Gene expression consists of transcription of DNA into mRNA (top) and the translation (center) of mRNA (with the help of a ribosome) into a protein (bottom).

in DNA and specifies the insertion of a specific amino acid into a protein. Proteins (lower part of Figure 1.8) are polymers made up of amino acid monomers. There are 20 different amino acids commonly found in proteins.

Protein assembly is accomplished with the aid of adapter molecules called **transfer RNA (tRNA)**. Within the ribosome, tRNAs recognize the information encoded in the mRNA codons and carry the proper amino acids for construction of the protein during translation.

We now know that gene expression can be more complex than outlined here. Some of these complexities will be discussed later in the text (see Chapters 14 and 19).

Proteins and Biological Function

In most cases, proteins are the end products of gene expression. The diversity of proteins and the biological functions they perform—the diversity of life itself—arises from the fact that proteins are made from combinations of 20 different amino acids. Consider that a protein chain containing 100 amino acids can have at each position any one of 20 amino acids; the number of possible different 100-amino-acid proteins, each with a unique sequence, is therefore equal to

$$20^{100}$$

Obviously, proteins are molecules with the potential for enormous structural diversity and serve as the mainstay of biological systems.

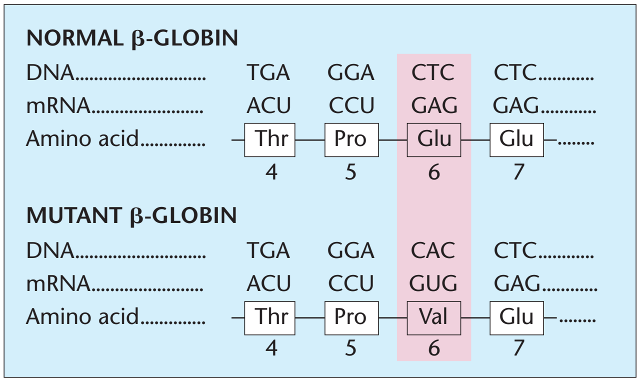
Enzymes form the largest category of proteins. These molecules serve as biological catalysts, lowering the energy of activation in reactions and allowing cellular metabolism to proceed at body temperature.

Proteins other than enzymes are critical components of cells and organisms. These include hemoglobin, the oxygen-binding molecule in red blood cells; insulin, a pancreatic hormone; collagen, a connective tissue molecule; and actin and myosin, the contractile muscle proteins. A protein's shape and chemical behavior are determined by its linear sequence of amino acids, which in turn is dictated by the stored information in the DNA of a gene that is transferred to RNA, which then directs the protein's synthesis.

Linking Genotype to Phenotype: Sickle-Cell Anemia

Once a protein is made, its biochemical or structural properties play a role in producing a phenotype. When mutation alters a gene, it may modify or even eliminate the encoded protein's usual function and cause an altered phenotype. To trace this chain of events, we will examine sickle-cell anemia, a human genetic disorder.

Sickle-cell anemia is caused by a mutant form of hemoglobin, the protein that transports oxygen from the lungs to cells in the body. Hemoglobin is a composite molecule made up of two different proteins, α-globin and β-globin, each encoded by a different gene. In sickle-cell anemia, a mutation in the gene encoding β-globin causes an amino acid substitution in 1 of the 146 amino acids in the protein. **Figure 1.9** shows the DNA sequence, the corresponding mRNA codons, and the amino acids occupying positions 4–7 for the normal and mutant forms of β-globin. Notice that the mutation in sickle-cell anemia consists of a change in one DNA nucleotide, which leads to a change in codon 6 in mRNA from GAG to GUG, which in turn changes amino acid number 6 in β-globin from glutamic acid to valine. The other 145 amino acids in the protein are not changed by this mutation.



NORMAL β-GLOBIN

DNA..........	TGA	GGA	CTC	CTC..........
mRNA..........	ACU	CCU	GAG	GAG..........
Amino acid..........	Thr	Pro	Glu	Glu..........
	4	5	6	7

MUTANT β-GLOBIN

DNA..........	TGA	GGA	CAC	CTC..........
mRNA..........	ACU	CCU	GUG	GAG..........
Amino acid..........	Thr	Pro	Val	Glu..........
	4	5	6	7

FIGURE 1.9 A single-nucleotide change in the DNA encoding β-globin (CTC → CAC) leads to an altered mRNA codon (GAG → GUG) and the insertion of a different amino acid (Glu → Val), producing the altered version of the β-globin protein that is responsible for sickle-cell anemia.

Individuals with two mutant copies of the β-globin gene have sickle-cell anemia. Their mutant β-globin proteins cause hemoglobin molecules in red blood cells to polymerize when the blood's oxygen concentration is low, forming long chains of hemoglobin that distort the shape of red blood cells (**Figure 1.10**). The deformed cells are fragile and break

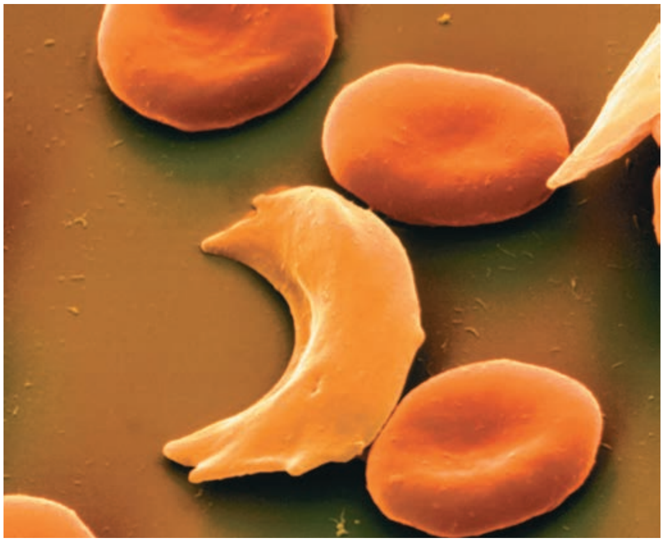


FIGURE 1.10 Normal red blood cells (round) and sickled red blood cells. The sickled cells block capillaries and small blood vessels.

easily, reducing the number of red blood cells in circulation (anemia is an insufficiency of red blood cells). Sickle-shaped blood cells block blood flow in capillaries and small blood vessels, causing severe pain and damage to the heart, brain, muscles, and kidneys. All the symptoms of this disorder are caused by a change in a single nucleotide in a gene that changes one amino acid out of 146 in the β-globin molecule, demonstrating the close relationship between genotype and phenotype.

1.4 Development of Recombinant DNA Technology Began the Era of DNA Cloning

The era of recombinant DNA began in the early 1970s, when researchers discovered that **restriction enzymes**, used by bacteria to cut and inactivate the DNA of invading viruses, could be used to cut any organism's DNA at specific nucleotide sequences, producing a reproducible set of fragments.

Soon after, researchers discovered ways to insert the DNA fragments produced by the action of restriction enzymes into carrier DNA molecules called **vectors** to form recombinant DNA molecules. When transferred into bacterial cells, thousands of copies, or **clones**, of the combined vector and DNA fragments are produced during bacterial reproduction. Large amounts of cloned DNA fragments can be isolated from these bacterial host cells. These DNA fragments can be used to isolate genes, to study their organization and expression, and to study their nucleotide sequence and evolution.

Collections of clones that represent an organism's **genome**, defined as the complete haploid DNA content of a specific organism, are called genomic libraries. Genomic libraries are now available for hundreds of species.

Recombinant DNA technology has not only accelerated the pace of research but also given rise to the biotechnology industry, which has grown to become a major contributor to the U.S. economy.

1.5 The Impact of Biotechnology Is Continually Expanding

The use of recombinant DNA technology and other molecular techniques to make products is called **biotechnology**. In the United States, biotechnology has quietly revolutionized many aspects of everyday life; products made by biotechnology are now found in the supermarket, in health care, in agriculture, and in the court system. A later chapter (see Chapter 22) contains a detailed discussion of biotechnology, but for now, let's look at some everyday examples of biotechnology's impact.

Plants, Animals, and the Food Supply

The use of recombinant DNA technology to genetically modify crop plants has revolutionized agriculture. Genes for traits including resistance to herbicides, insects, and genes for nutritional enhancement have been introduced into crop plants. The transfer of heritable traits across species using recombinant DNA technology creates **transgenic** organisms. Herbicide-resistant corn and soybeans were first planted in the mid-1990s, and transgenic strains now represent about 88 percent of the U.S. corn crop and 93 percent of the U.S. soybean crop. It is estimated that more than 70 percent of the processed food in the United States contains ingredients from transgenic crops.

We will discuss the most recent findings involving genetically modified organisms later in the text. (Special Topics Chapter 4—Genetically Modified Foods).

New methods of cloning livestock such as sheep and cattle have also changed the way we use these animals. In 1996, Dolly the sheep (**Figure 1.11**) was cloned by nuclear transfer, a method in which the nucleus of an adult cell is transferred into an egg that has had its nucleus removed. This method makes it possible to produce dozens or hundreds of genetically identical offspring with desirable traits and has many applications in agriculture, sports, and medicine.

Biotechnology has also changed the way human proteins for medical use are produced. Through use of gene transfer, transgenic animals now synthesize these therapeutic

FIGURE 1.11 Dolly, a Finn Dorset sheep cloned from the genetic material of an adult mammary cell, shown next to her first-born lamb, Bonnie.

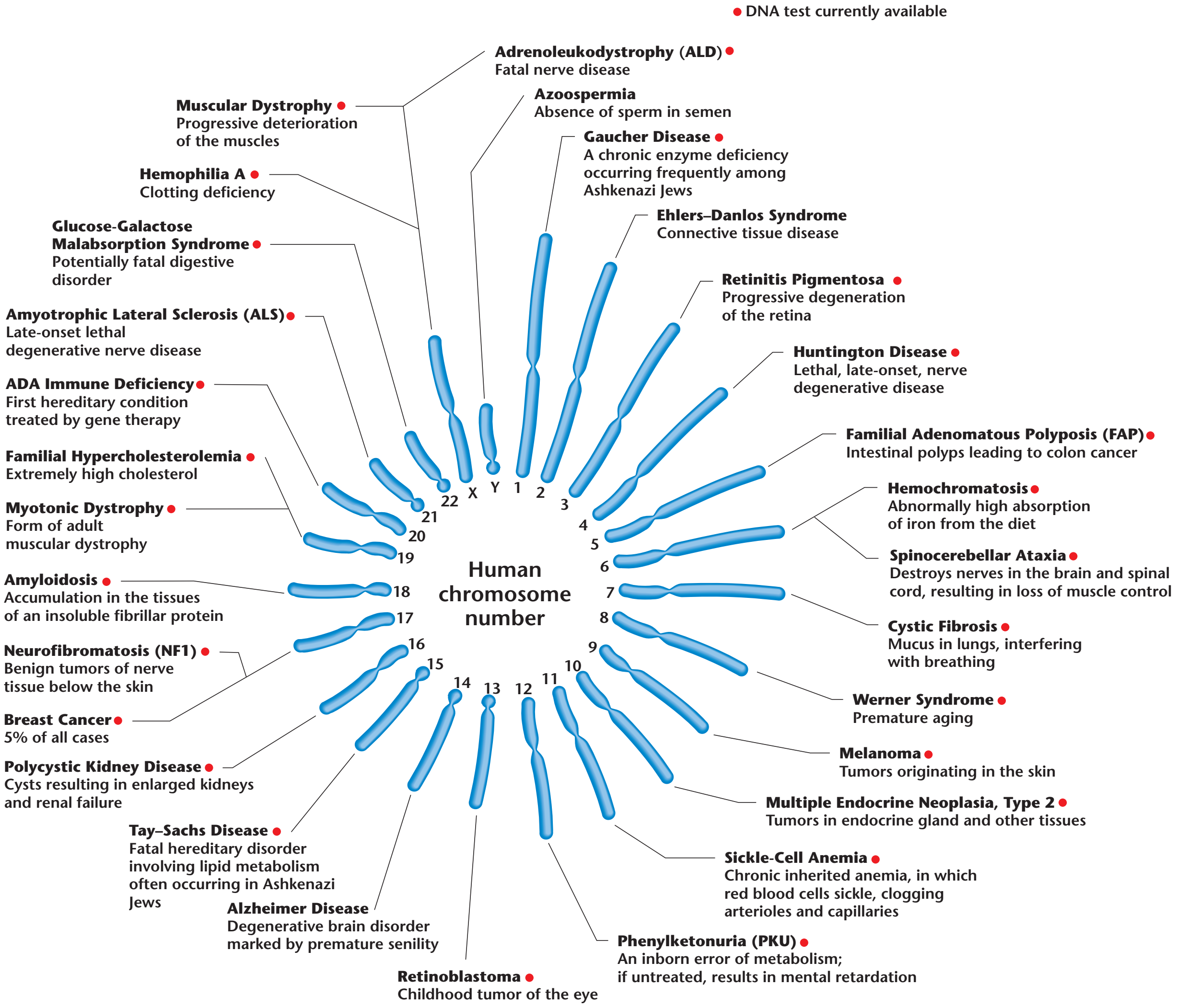

● DNA test currently available
Adrenoleukodystrophy (ALD) ●
Fatal nerve disease
Muscular Dystrophy ●
Progressive deterioration of the muscles
Azoospermia
Absence of sperm in semen
Gaucher Disease ●
A chronic enzyme deficiency occurring frequently among Ashkenazi Jews
Hemophilia A ●
Clotting deficiency
Glucose-Galactose Malabsorption Syndrome ●
Potentially fatal digestive disorder
Ehlers–Danlos Syndrome
Connective tissue disease
Retinitis Pigmentosa ●
Progressive degeneration of the retina
Amyotrophic Lateral Sclerosis (ALS) ●
Late-onset lethal degenerative nerve disease
Huntington Disease ●
Lethal, late-onset, nerve degenerative disease
ADA Immune Deficiency ●
First hereditary condition treated by gene therapy
Familial Adenomatous Polyposis (FAP) ●
Intestinal polyps leading to colon cancer
Familial Hypercholesterolemia ●
Extremely high cholesterol
Hemochromatosis ●
Abnormally high absorption of iron from the diet
Myotonic Dystrophy ●
Form of adult muscular dystrophy
Spinocerebellar Ataxia ●
Destroys nerves in the brain and spinal cord, resulting in loss of muscle control
Amyloidosis ●
Accumulation in the tissues of an insoluble fibrillar protein
Human chromosome number
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Cystic Fibrosis ●
Mucus in lungs, interfering with breathing
Neurofibromatosis (NF1) ●
Benign tumors of nerve tissue below the skin
Werner Syndrome ●
Premature aging
Breast Cancer ●
5% of all cases
Melanoma ●
Tumors originating in the skin
Polycystic Kidney Disease ●
Cysts resulting in enlarged kidneys and renal failure
Multiple Endocrine Neoplasia, Type 2 ●
Tumors in endocrine gland and other tissues
Tay–Sachs Disease ●
Fatal hereditary disorder involving lipid metabolism often occurring in Ashkenazi Jews
Sickle-Cell Anemia ●
Chronic inherited anemia, in which red blood cells sickle, clogging arterioles and capillaries
Alzheimer Disease
Degenerative brain disorder marked by premature senility
Phenylketonuria (PKU) ●
An inborn error of metabolism; if untreated, results in mental retardation
Retinoblastoma ●
Childhood tumor of the eye

FIGURE 1.12 The human chromosome set, showing the location of some genes whose mutant forms cause hereditary diseases. Conditions that can be diagnosed using genetic testing are indicated by a red dot.

proteins. In 2009, an anticlotting protein derived from the milk of transgenic goats was approved by the U.S. Food and Drug Administration for use in the United States. Other human proteins from transgenic animals are now being used in clinical trials to treat several diseases. The biotechnology revolution will continue to expand as new methods are developed to make an increasing array of products.

Biotechnology in Genetics and Medicine

More than 10 million children or adults in the United States suffer from some form of genetic disorder, and every childbearing couple faces an approximately 3 percent risk of having a child with a genetic anomaly. The molecular basis for hundreds of genetic disorders is now known, and many of these genes have been mapped, isolated, and cloned (Figure 1.12). Biotechnology-derived genetic testing is now available to perform prenatal diagnosis of heritable disorders and to test parents for their status as "carriers" of more than 100 inherited disorders. Newer methods now under development offer the possibility of scanning an entire genome to establish an individual's risk of developing a genetic disorder or having an affected child. The use of genetic testing and related technologies raises ethical concerns that have yet to be resolved.

1.6 Genomics, Proteomics, and Bioinformatics Are New and Expanding Fields

The use of recombinant DNA technology to create genomic libraries prompted scientists to consider sequencing all the clones in a library to derive the nucleotide sequence of an organism's genome. This sequence information would be used to identify each gene in the genome and establish its function.

One such project, the Human Genome Project, began in 1990 as an international effort to sequence the human genome. By 2003, the publicly funded Human Genome Project and a private, industry-funded genome project completed sequencing of the gene-containing portion of the genome.

As more genome sequences were acquired, several new biological disciplines arose. One, called **genomics** (the study of genomes), studies the structure, function, and evolution of genes and genomes. A second field, **proteomics**, identifies the set of proteins present in a cell under a given set of conditions, and studies their functions and interactions. To store, retrieve, and analyze the massive amount of data generated by genomics and proteomics, a specialized subfield of information technology called **bioinformatics** was created to develop hardware and software for processing nucleotide and protein data.

Geneticists and other biologists now use information in databases containing nucleic acid sequences, protein sequences, and gene-interaction networks to answer experimental questions in a matter of minutes instead of months and years. A feature called "Exploring Genomics," located at the end of many of the chapters in this textbook, gives you the opportunity to explore these databases for yourself while completing an interactive genetics exercise.

Modern Approaches to Understanding Gene Function

This edition continues the feature "Modern Approaches to Understanding Gene Function" that appears in selected chapters. It is designed to introduce you to examples of the most current experimental approaches used by geneticists to study gene function. Its placement within these chapters links the techniques to the concepts that have just been presented.

Historically, an approach referred to as **classical** or **forward genetics** was essential for studying and understanding gene function. In this approach geneticists relied on the use of naturally occurring mutations or intentionally induced mutations (using chemicals, X-rays or UV light as examples) to cause altered phenotypes in model organisms, and then worked through the lab-intensive and time-consuming process of identifying the genes that caused these new phenotypes. Such characterization often led to the identification of the gene or genes of interest, and once the technology advanced, the gene sequence could be determined.

Classical genetics approaches are still used, but as whole genome sequencing has become routine, molecular approaches to understanding gene function have changed considerably in genetic research. These modern approaches are what we will highlight in this feature.

For the past two decades or so, geneticists have relied on the use of molecular techniques incorporating an approach referred to as **reverse genetics**. In reverse genetics, the DNA sequence for a particular gene of interest is known, but the role and function of the gene are typically not well understood. For example, molecular biology techniques such as **gene knockout** render targeted genes nonfunctional in a model organism or in cultured cells, allowing scientists to investigate the fundamental question of "what happens if this gene is disrupted?" After making a knockout organism, scientists look for both apparent phenotype changes, as well as those at the cellular and molecular level. The ultimate goal is to determine the function of the gene.

In "Modern Approaches to Understanding Gene Function" we will highlight experimental examples of how gene function has been revealed through modern applications of molecular techniques involving reverse genetics. You will learn about gene knockouts, transgenic animals, transposon-mediated mutagenesis, gene overexpression, and RNA interference-based methods for interrupting genes, among other approaches. Our hope is to bring you to the "cutting edge" of genetic studies.

1.7 Genetic Studies Rely on the Use of Model Organisms

After the rediscovery of Mendel's work in 1900, research using a wide range of organisms confirmed that the principles of inheritance he described were of universal significance among plants and animals. Geneticists gradually came to focus attention on a small number of organisms, including the fruit fly (*Drosophila melanogaster*) and the mouse (*Mus musculus*) (**Figure 1.13**). This trend developed for two main reasons: first, it was clear that genetic mechanisms were the same in most organisms, and second, these organisms had characteristics that made them especially suitable for genetic research. They were easy to grow, had relatively

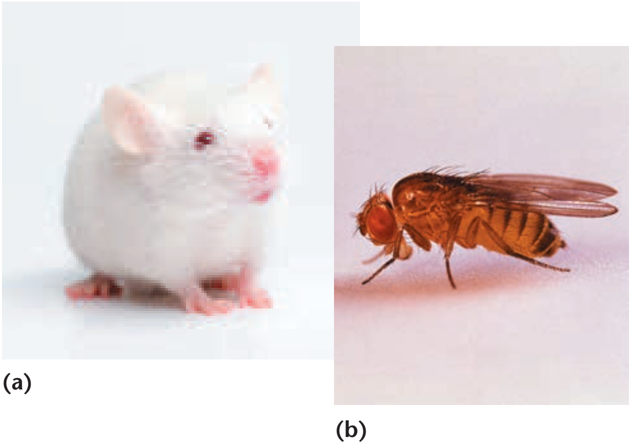


FIGURE 1.13 The first generation of model organisms in genetic analysis included (a) the mouse, *Mus musculus* and (b) the fruit fly, *Drosophila melanogaster*.

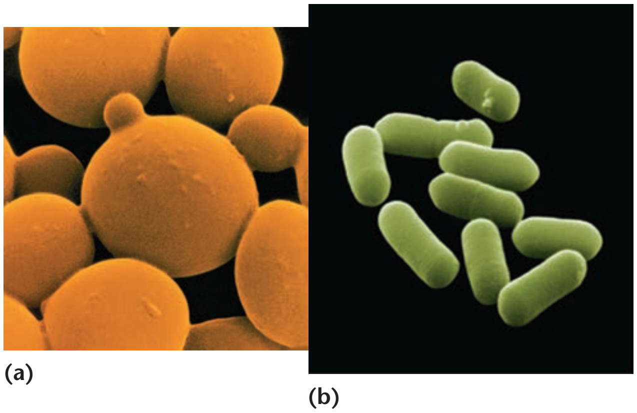


FIGURE 1.14 Microbes that have become model organisms for genetic studies include (a) the yeast *Saccharomyces cerevisiae* and (b) the bacterium *Escherichia coli*.

short life cycles, produced many offspring, and their genetic analysis was fairly straightforward. Over time, researchers created a large catalog of mutant strains for these species, and the mutations were carefully studied, characterized, and mapped. Because of their well-characterized genetics, these species became **model organisms**, defined as organisms used for the study of basic biological processes. In later chapters, we will see how discoveries in model organisms are shedding light on many aspects of biology, including aging, cancer, the immune system, and behavior.

The Modern Set of Genetic Model Organisms

Gradually, geneticists added other species to their collection of model organisms: viruses (such as the T phages and lambda phage) and microorganisms (the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae*) (**Figure 1.14**).

More recently, additional species have been developed as model organisms, three of which are shown in the chapter opening photograph. Each species was chosen to allow study of some aspect of embryonic development. The nematode *Caenorhabditis elegans* was chosen as a model system to study the development and function of the nervous system because its nervous system contains only a few hundred cells and the developmental fate of these and all other cells in the body has been mapped out. *Arabidopsis thaliana*, a small plant with a short life cycle, has become a model organism for the study of many aspects of plant biology. The zebrafish, *Danio rerio*, is used to study vertebrate development: it is small, it reproduces rapidly, and its egg, embryo, and larvae are all transparent.

Model Organisms and Human Diseases

The development of recombinant DNA technology and the results of genome sequencing have confirmed that all life has a common origin. Because of this, genes with similar functions in different organisms tend to be similar or identical in structure and nucleotide sequence. Much of what scientists learn by studying the genetics of model organisms can therefore be applied to humans as the basis for understanding and treating human diseases. In addition, the ability to create transgenic organisms by transferring genes between species has enabled scientists to develop models of human diseases in organisms ranging from bacteria to fungi, plants, and animals (**Table 1.1**).

TABLE 1.1 Model Organisms Used to Study Some Human Diseases

Organism	Human Diseases
E. coli	Colon cancer and other cancers
S. cerevisiae	Cancer, Werner syndrome
D. melanogaster	Disorders of the nervous system, cancer
C. elegans	Diabetes
D. rerio	Cardiovascular disease
M. musculus	Lesch–Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases

The idea of studying a human disease such as colon cancer by using *E. coli* may strike you as strange, but the basic steps of DNA repair (a process that is defective in some forms of colon cancer) are the same in both organisms, and a gene involved in DNA repair (*mutL* in *E. coli* and *MLH1* in humans) is found in both organisms. More importantly, *E. coli* has the advantage of being easier to grow (the cells divide every 20 minutes), and researchers can easily create and study new mutations in the bacterial *mutL* gene in order to figure out how it works. This knowledge may eventually lead to the development of drugs and other therapies to treat colon cancer in humans.

The fruit fly, *Drosophila melanogaster*, is also being used to study a number of human diseases. Mutant genes have been identified in *D. melanogaster* that produce phenotypes with structural abnormalities of the nervous system and adult-onset degeneration of the nervous system. The information from genome-sequencing projects indicates that almost all these genes have human counterparts. For example, genes involved in a complex human disease of the retina called retinitis pigmentosa are identical to *Drosophila* genes involved in retinal degeneration. Study of these mutations in *Drosophila* is helping to dissect this complex disease and identify the function of the genes involved.

Another approach to studying diseases of the human nervous system is to transfer mutant human disease genes into *Drosophila* using recombinant DNA technology. The transgenic flies are then used for studying the mutant human genes themselves, other genes that affect the expression of the human disease genes, and the effects of therapeutic drugs on the action of those genes—all studies that are difficult or impossible to perform in humans. This gene transfer approach is being used to study almost a dozen human neurodegenerative disorders, including Huntington disease, Machado–Joseph disease, myotonic dystrophy, and Alzheimer disease.

Throughout the following chapters, you will encounter these model organisms again and again. Remember each time you meet them that they not only have a rich history in basic genetics research but are also at the forefront in the study of human genetic disorders and infectious diseases. As discussed in the next section, however, we have yet to reach a consensus on how and when some of this technology will be accepted as safe and ethically acceptable.

1.8 We Live in the Age of Genetics

Mendel described his decade-long project on inheritance in pea plants in an 1865 paper presented at a meeting of the Natural History Society of Brünn in Moravia. Less than 100 years later, the 1962 Nobel Prize was awarded to James Watson, Francis Crick, and Maurice Wilkins for their work on the structure of DNA. This time span encompassed the years leading up to the acceptance of Mendel's work, the discovery that genes are on chromosomes, the experiments that proved DNA encodes genetic information, and the elucidation of the molecular basis for DNA replication. The rapid development of genetics from Mendel's monastery garden to the Human Genome Project and beyond is summarized in a timeline in **Figure 1.15**.

The Nobel Prize and Genetics

No other scientific discipline has experienced the explosion of information and the level of excitement generated by the discoveries in genetics. This impact is especially apparent in the list of Nobel Prizes related to genetics, beginning with those awarded in the early and mid-twentieth century and continuing into the present (see inside front cover). Nobel Prizes in Medicine or Physiology and Chemistry have been consistently awarded for work in genetics and related fields. One of the first such prizes awarded was given to Thomas H. Morgan in 1933 for his research on the chromosome theory of inheritance. That award was followed by many others, including prizes for the discovery of genetic recombination, the relationship between genes and proteins, the structure of DNA, and the genetic code. This trend has continued throughout the twentieth and twenty-first centuries. The advent of genomic studies and the applications of such findings will most certainly lead the way for future awards.

Genetics, Ethics, and Society

Just as there has never been a more exciting time to study genetics, the impact of this discipline on society has never been more profound. Genetics and its applications in biotechnology are developing much faster than the social conventions, public policies, and laws required to regulate their use. As a society, we are grappling with a host of sensitive genetics-related issues, including concerns about prenatal testing, genetic discrimination, ownership of genes, access to and safety of gene therapy, and genetic privacy. Two features appearing at the end of most chapters, "Case Study" and "Genetics, Ethics, and Society," consider ethical issues raised by the use of genetic technology. This emphasis on ethics reflects the growing concern and dilemmas that advances in genetics pose to our society and the future of our species.

By the time you finish this course, you will have seen more than enough evidence to convince yourself that the present is the Age of Genetics, and you will understand the need to think about and become a participant in the dialogue concerning genetic science and its use.

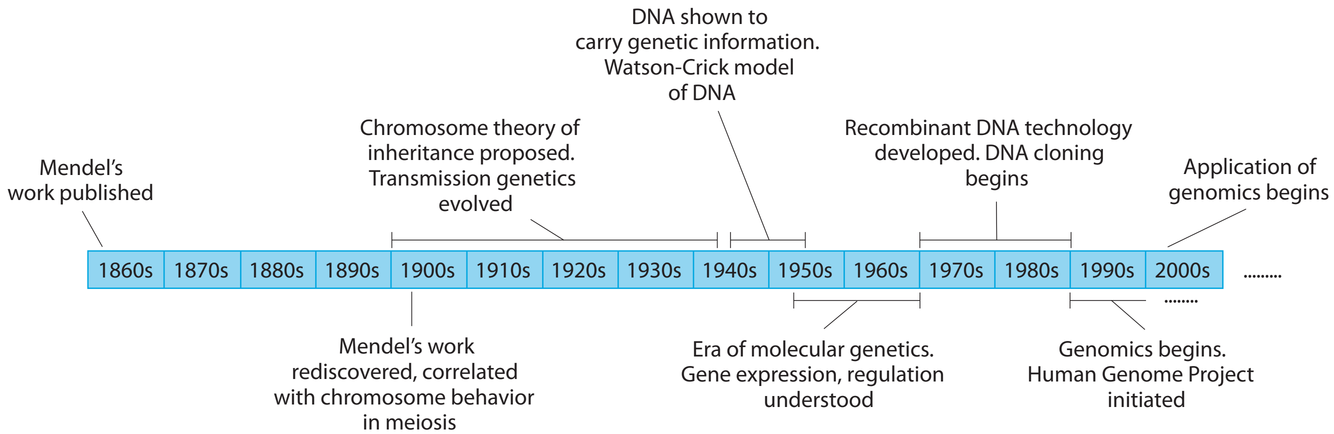

DNA shown to carry genetic information. Watson-Crick model of DNA
Chromosome theory of inheritance proposed. Transmission genetics evolved
Recombinant DNA technology developed. DNA cloning begins
Mendel's work published
Application of genomics begins
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
Mendel's work rediscovered, correlated with chromosome behavior in meiosis
Era of molecular genetics. Gene expression, regulation understood
Genomics begins. Human Genome Project initiated

FIGURE 1.15 A timeline showing the development of genetics from Gregor Mendel's work on pea plants to the current era of genomics and its many applications in research, medicine, and society. Having a sense of the history of discovery in genetics should provide you with a useful framework as you proceed through this textbook.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Mendel's work on pea plants established the principles of gene transmission from parents to offspring that form the foundation for the science of genetics.
2. Genes and chromosomes are the fundamental units in the chromosomal theory of inheritance. This theory explains that inherited traits are controlled by genes located on chromosomes and shows how the transmission of genetic information maintains genetic continuity from generation to generation.
3. Molecular genetics—based on the central dogma that DNA is a template for making RNA, which encodes the order of amino acids in proteins—explains the phenomena described by Mendelian genetics, referred to as transmission genetics.
4. Recombinant DNA technology, a far-reaching methodology used in molecular genetics, allows genes from one organism to be spliced into vectors and cloned, producing many copies of specific DNA sequences.
5. Biotechnology has revolutionized agriculture, the pharmaceutical industry, and medicine. It has made possible the mass production of medically important gene products. Genetic testing allows detection of individuals with genetic disorders and those at risk of having affected children, and gene therapy offers hope for the treatment of serious genetic disorders.
6. Genomics, proteomics, and bioinformatics are new fields derived from recombinant DNA technology. These fields combine genetics with information technology and allow scientists to explore genome sequences, the structure and function of genes, the protein set within cells, and the evolution of genomes. The Human Genome Project is one example of genomics.
7. The use of model organisms has advanced the understanding of genetic mechanisms and, coupled with recombinant DNA technology, has produced models of human genetic diseases.
8. The effects of genetic technology on society are profound, and the development of policy and legislation to deal with issues derived from the use of this technology is lagging behind the resulting innovations.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. Describe Mendel's conclusions about how traits are passed from generation to generation.

2. CONCEPT QUESTION Review the Chapter Concepts list on p. 1. Most of these are related to the discovery of DNA as the genetic material and the subsequent development of recombinant DNA technology. Write a brief essay that discusses the impact of recombinant DNA technology on genetics as we perceive the discipline today.

3. What is the chromosome theory of inheritance, and how is it related to Mendel's findings?

4. Define genotype and phenotype. Describe how they are related and how alleles fit into your definitions.

5. Given the state of knowledge at the time of the Avery, MacLeod, and McCarty experiment, why was it difficult for some scientists to accept that DNA is the carrier of genetic information?

6. Contrast chromosomes and genes.

7. How is genetic information encoded in a DNA molecule?

8. Describe the central dogma of molecular genetics and how it serves as the basis of modern genetics.

9. How many different proteins, each with a unique amino acid sequence, can be constructed that have a length of five amino acids?

10. Outline the roles played by restriction enzymes and vectors in cloning DNA.

11. What are some of the impacts of biotechnology on crop plants in the United States?

12. Summarize the arguments for and against patenting genetically modified organisms.

13. We all carry about 20,000 genes in our genome. So far, patents have been issued for more than 6000 of these genes. Do you think that companies or individuals should be able to patent human genes? Why or why not?

14. How has the use of model organisms advanced our knowledge of the genes that control human diseases?

15. If you knew that a devastating late-onset inherited disease runs in your family (in other words, a disease that does not appear until later in life) and you could be tested for it at the age of 20, would you want to know whether you are a carrier? Would your answer be likely to change when you reach age 40?

16. Why do you think discoveries in genetics have been recognized with so many Nobel Prizes?

17. The Age of Genetics was created by remarkable advances in the use of biotechnology to manipulate plant and animal genomes. Given that the world population reached 7.5 billion people in 2017 and is expected to reach 9.7 billion in 2050, some scientists have proposed that only the worldwide introduction of genetically modified (GM) foods will increase crop yields enough to meet future nutritional demands. Pest resistance, herbicide, cold, drought, and salinity tolerance, along with increased nutrition, are seen as positive attributes of GM foods. However, others caution that unintended harm to other organisms, reduced effectiveness to pesticides, gene transfer to nontarget species, allergenicity, and as yet unknown effects on human health are potential concerns regarding GM foods. If you were in a position to control the introduction of a GM primary food product (rice, for example), what criteria would you establish before allowing such introduction?

2

Mitosis and Meiosis

Chromosomes in the prometaphase stage of mitosis, derived from a cell in the flower of *Haemanthus*.

CHAPTER CONCEPTS

- Genetic continuity between generations of cells and between generations of sexually reproducing organisms is maintained through the processes of mitosis and meiosis, respectively.
- Diploid eukaryotic cells contain their genetic information in pairs of homologous chromosomes, with one member of each pair being derived from the maternal parent and one from the paternal parent.
- Mitosis provides a mechanism by which chromosomes, having been duplicated, are distributed into progeny cells during cell reproduction.
- Mitosis converts a diploid cell into two diploid daughter cells.
- The process of meiosis distributes one member of each homologous pair of chromosomes into each gamete or spore, thus reducing the diploid chromosome number to the haploid chromosome number.
- Meiosis generates genetic variability by distributing various combinations of maternal and paternal members of each homologous pair of chromosomes into gametes or spores.
- During the stages of mitosis and meiosis, the genetic material is condensed into discrete structures called chromosomes.

Every living thing contains a substance described as the genetic material. Except in certain viruses, this material is composed of the nucleic acid DNA. DNA has an underlying linear structure possessing segments called genes, the products of which direct the metabolic activities of cells. An organism's DNA, with its arrays of genes, is organized into structures called **chromosomes**, which serve as vehicles for transmitting genetic information. The manner in which chromosomes are transmitted from one generation of cells to the next and from organisms to their descendants must be exceedingly precise. In this chapter we consider exactly how genetic continuity is maintained between cells and organisms.

Two major processes are involved in the genetic continuity of nucleated cells: **mitosis** and **meiosis**. Although the mechanisms of the two processes are similar in many ways, the outcomes are quite different. Mitosis leads to the production of two cells, each with the same number of chromosomes as the parent cell. In contrast, meiosis reduces the genetic content and the number of chromosomes by precisely half. This reduction is essential if sexual reproduction is to occur without doubling the amount of genetic material in each new generation. Strictly speaking, mitosis is that portion of the cell cycle during which the hereditary components are equally partitioned into daughter cells. Meiosis is part of a special type of cell division that leads to the production of sex cells: **gametes** or **spores**. This process is an essential step in the transmission of genetic information from an organism to its offspring.

Normally, chromosomes are visible only during mitosis and meiosis. When cells are not undergoing division, the genetic material making up chromosomes unfolds and uncoils into a diffuse network within the nucleus, generally referred to as **chromatin**. Before describing mitosis and meiosis, we will briefly review the structure of cells, emphasizing components that are of particular significance to genetic function. We will also compare the structural differences between the prokaryotic (nonnucleated) cells of bacteria and the eukaryotic cells of higher organisms. We then devote the remainder of the chapter to the behavior of chromosomes during cell division.

2.1 Cell Structure Is Closely Tied to Genetic Function

Before 1940, our knowledge of cell structure was limited to what we could see with the light microscope. Around 1940, the transmission electron microscope was in its early stages of development, and by 1950, many details of cell ultrastructure had emerged. Under the electron microscope, cells were seen as highly varied, highly organized structures whose form and function are dependent on specific genetic expression by each cell type. A new world of whorled membranes, organelles, microtubules, granules, and filaments was revealed. These discoveries revolutionized thinking in the entire field of biology. Many cell components, such as the nucleolus, ribosome, and centriole, are involved directly or indirectly with genetic processes. Other components—the mitochondria and chloroplasts—contain their own unique genetic information. Here, we will focus primarily on those aspects of cell structure that relate to genetic study. The generalized animal cell shown in Figure 2.1 illustrates most of the structures we will discuss.

All cells are surrounded by a ***plasma membrane,*** an outer covering that defines the cell boundary and delimits the cell from its immediate external environment. This membrane is not passive but instead actively controls the movement of materials into and out of the cell. In addition to this membrane, plant cells have an outer covering called the ***cell wall*** whose major component is a polysaccharide called *cellulose*.

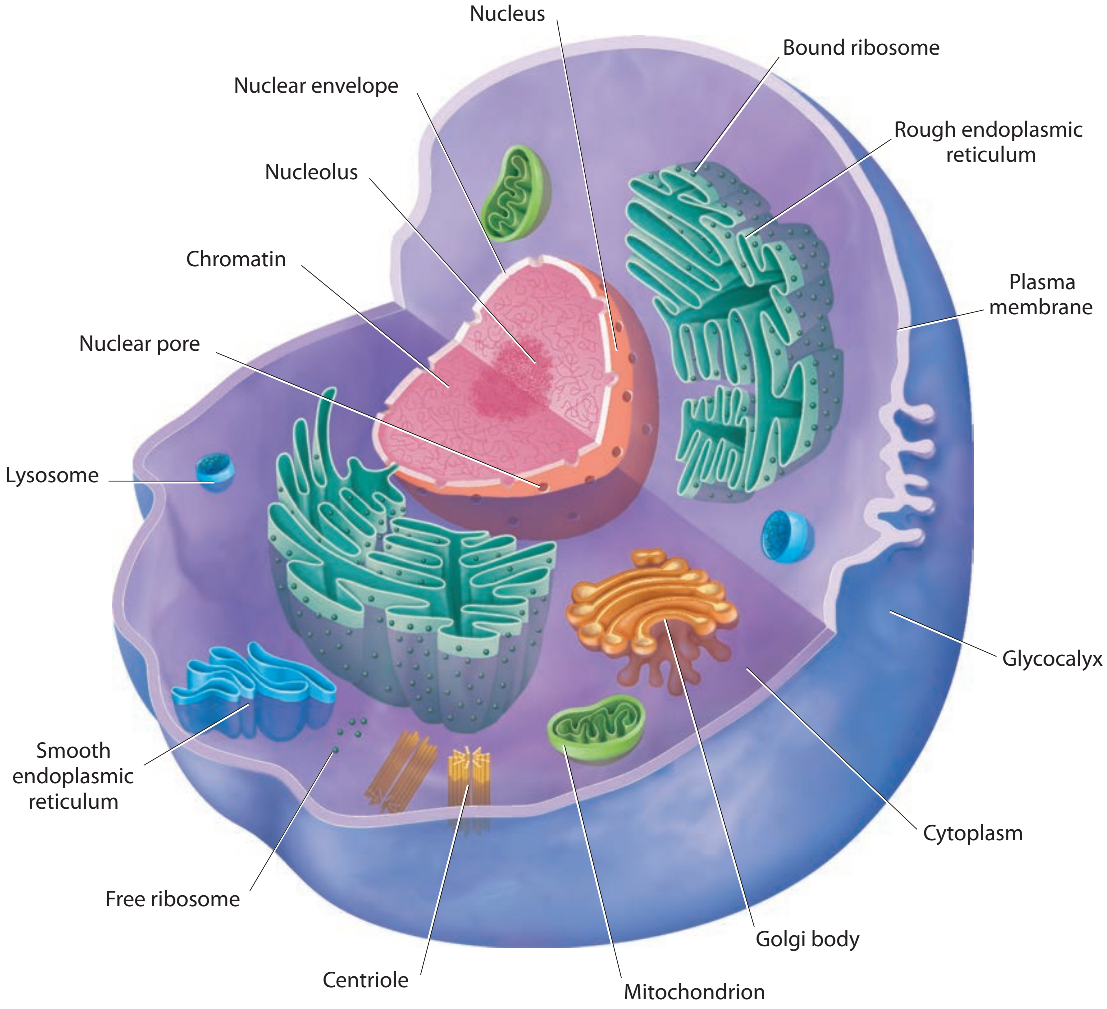


FIGURE 2.1 A generalized animal cell. The cellular components discussed in the text are emphasized here.

Many, if not most, animal cells have a covering over the plasma membrane, referred to as the **glycocalyx**, or ***cell coat***. Consisting of glycoproteins and polysaccharides, this covering has a chemical composition that differs from comparable structures in either plants or bacteria. The glycocalyx, among other functions, provides biochemical identity at the surface of cells, and the components of the coat that establish cellular identity are under genetic control. For example, various cell-identity markers that you may have heard of—the AB, Rh, and MN antigens—are found on the surface of red blood cells, among other cell types. On the surface of other cells, histocompatibility antigens, which elicit an immune response during tissue and organ transplants, are present. Various **receptor molecules** are also found on the surfaces of cells. These molecules act as recognition sites that transfer specific chemical signals across the cell membrane into the cell.

Living organisms are categorized into two major groups depending on whether or not their cells contain a nucleus. The presence of a nucleus and other membranous organelles is the defining characteristic of **eukaryotic organisms**. The **nucleus** in eukaryotic cells is a membrane-bound structure that houses the genetic material, DNA, which is complexed with an array of acidic and basic proteins into thin fibers. During nondivisional phases of the cell cycle, the fibers are uncoiled and dispersed into chromatin (as mentioned above). During mitosis and meiosis, chromatin fibers coil and condense into chromosomes. Also present in the nucleus is the **nucleolus**, an amorphous component where ribosomal RNA (rRNA) is synthesized and where the initial stages of ribosomal assembly occur. The portions of DNA that encode rRNA are collectively referred to as the **nucleolus organizer region**, or the **NOR**.

Prokaryotic organisms, of which there are two major groups, lack a nuclear envelope and membranous organelles. For the purpose of our brief discussion here, we will consider the *eubacteria,* the other group being the more ancient bacteria referred to as *archaea.* In eubacteria, such as *Escherichia coli,* the genetic material is present as a long, circular DNA molecule that is compacted into an unenclosed region called the **nucleoid**. Part of the DNA may be attached to the cell membrane, but in general the nucleoid extends through a large part of the cell. Although the DNA is compacted, it does not undergo the extensive coiling characteristic of the stages of mitosis, during which the chromosomes of eukaryotes become visible. Nor is the DNA associated as extensively with proteins as is eukaryotic DNA. Figure 2.2, which shows two bacteria forming by cell division, illustrates the nucleoid regions containing the bacterial chromosomes. Prokaryotic cells do not have a distinct nucleolus but do contain genes that specify rRNA molecules.

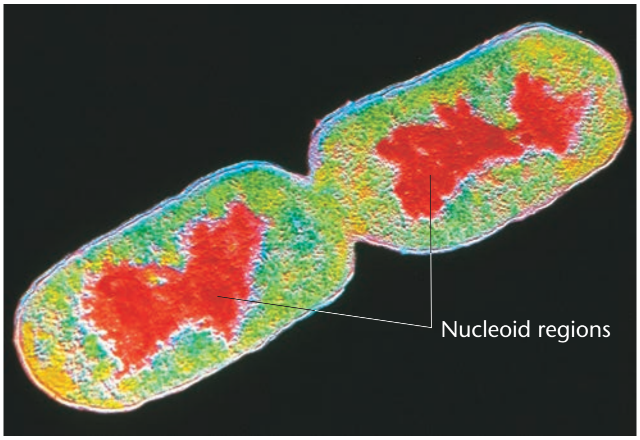


FIGURE 2.2 Color-enhanced electron micrograph of *E. coli* undergoing cell division. Particularly prominent are the two chromosomal areas (shown in red), called nucleoids, that have been partitioned into the daughter cells.

The remainder of the eukaryotic cell within the plasma membrane, excluding the nucleus, is referred to as **cytoplasm** and includes a variety of extranuclear cellular organelles. In the cytoplasm, a nonparticulate, colloidal material referred to as the ***cytosol*** surrounds and encompasses the cellular organelles. The cytoplasm also includes an extensive system of tubules and filaments, comprising the cytoskeleton, which provides a lattice of support structures within the cell. Consisting primarily of **microtubules**, which are made of the protein **tubulin**, and **microfilaments**, which derive from the protein ***actin***, this structural framework maintains cell shape, facilitates cell mobility, and anchors the various organelles.

One organelle, the membranous **endoplasmic reticulum (ER)**, compartmentalizes the cytoplasm, greatly increasing the surface area available for biochemical synthesis. The ER appears smooth in places where it serves as the site for synthesizing fatty acids and phospholipids; in other places, it appears rough because it is studded with ribosomes. **Ribosomes** serve as sites where genetic information contained in messenger RNA (mRNA) is translated into proteins.

Three other cytoplasmic structures are very important in the eukaryotic cell's activities: mitochondria, chloroplasts, and centrioles. **Mitochondria** are found in most eukaryotes, including both animal and plant cells, and are the sites of the oxidative phases of cell respiration. These chemical reactions generate large amounts of the energy-rich molecule adenosine triphosphate (ATP). **Chloroplasts**, which are found in plants, algae, and some protozoans, are associated with photosynthesis, the major energy-trapping process on Earth. Both mitochondria and chloroplasts contain DNA in a form distinct from that found in the nucleus. They are able to duplicate themselves and transcribe and translate their own genetic information.

Animal cells and some plant cells also contain a pair of complex structures called **centrioles**. These cytoplasmic bodies, each located in a specialized region called the **centrosome**, are associated with the organization of spindle fibers that function in mitosis and meiosis. In some organisms, the centriole is derived from another structure, the basal body, which is associated with the formation of cilia and flagella (hair-like and whip-like structures for propelling cells or moving materials).

The organization of **spindle fibers** by the centrioles occurs during the early phases of mitosis and meiosis. These fibers play an important role in the movement of chromosomes as they separate during cell division. They are composed of arrays of microtubules consisting of polymers of the protein tubulin.

2.2 Chromosomes Exist in Homologous Pairs in Diploid Organisms

As we discuss the processes of mitosis and meiosis, it is important that you understand the concept of homologous chromosomes. Such an understanding will also be of critical importance in our future discussions of Mendelian genetics. Chromosomes are most easily visualized during mitosis. When they are examined carefully, distinctive lengths and shapes are apparent. Each chromosome contains a constricted region called the **centromere**, whose location establishes the general appearance of each chromosome. Figure 2.3 shows chromosomes with centromere placements at different distances along their length. Extending from either side of the centromere are the arms of the chromosome. Depending on the position of the centromere, different arm ratios are produced. As Figure 2.3 illustrates, chromosomes are classified as **metacentric**, **submetacentric**, **acrocentric**, or **telocentric** on the basis of the centromere location. The shorter arm, by convention, is shown above the centromere and is called the **p arm** (p, for "petite"). The longer arm is shown below the centromere and is called the **q arm** (q because it is the next letter in the alphabet).

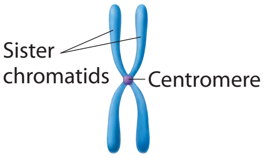
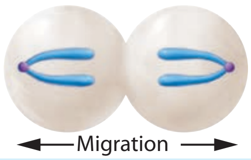
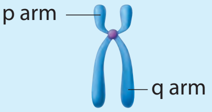
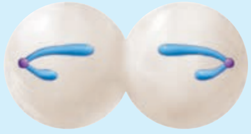
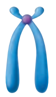
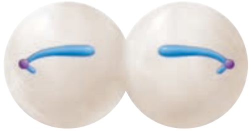
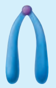
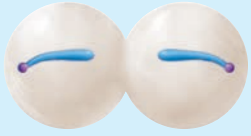
Centromere location	Designation	Metaphase shape	Anaphase shape
Middle	Metacentric		
Between middle and end	Submetacentric		
Close to end	Acrocentric		
At end	Telocentric		

FIGURE 2.3 Centromere locations and the chromosome designations that are based on them. Note that the shape of the chromosome during anaphase is determined by the position of the centromere during metaphase.

In the study of mitosis, several other observations are of particular relevance. First, all somatic cells derived from members of the same species contain an identical number of chromosomes. In most cases, this represents what is referred to as the **diploid number ($2n$).** When the lengths and centromere placements of all such chromosomes are examined, a second general feature is apparent. With the exception of sex chromosomes, they exist in pairs with regard to these two properties, and the members of each pair are called **homologous chromosomes**. So, for each chromosome exhibiting a specific length and centromere placement, another exists with identical features.

There are exceptions to this rule. Many bacteria and viruses have but one chromosome, and organisms such as yeasts and molds, and certain plants such as bryophytes (mosses), spend the predominant phase of their life cycle in the haploid stage. That is, they contain only one member of each homologous pair of chromosomes during most of their lives.

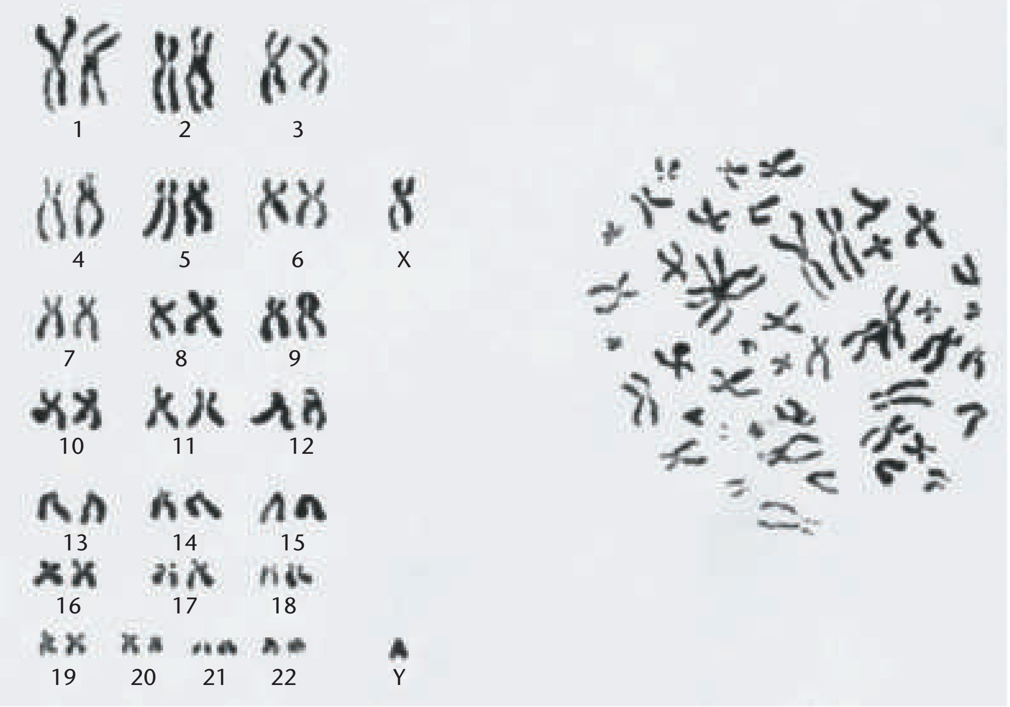


FIGURE 2.4 A metaphase preparation of chromosomes derived from a dividing cell of a human male (right), and the karyotype derived from the metaphase preparation (left). All but the X and Y chromosomes are present in homologous pairs. Each chromosome is clearly a double structure consisting of a pair of sister chromatids joined by a common centromere.

Figure 2.4 illustrates the physical appearance of different pairs of homologous chromosomes. There, the human mitotic chromosomes have been photographed, cut out of the print, and matched up, creating a display called a **karyotype**. As you can see, humans have a 2*n* number of 46 chromosomes, which on close examination exhibit a diversity of sizes and centromere placements. Note also that each of the 46 chromosomes in this karyotype is clearly a double structure consisting of two parallel *sister chromatids* connected by a common centromere. Had these chromosomes been allowed to continue dividing, the sister chromatids, which are replicas of one another, would have separated into the two new cells as division continued.

The **haploid number (*n*)** of chromosomes is equal to one-half the diploid number. Collectively, the genetic information contained in a haploid set of chromosomes constitutes the **genome** of the species. This, of course, includes copies of all genes as well as a large amount of noncoding DNA. The examples listed in Table 2.1 demonstrate the wide range of *n* values found in plants and animals.

Homologous chromosomes have important genetic similarities. They contain identical gene sites along their lengths; each site is called a **locus** (pl. loci). Thus, they are identical in the traits that they influence and in their genetic potential. In sexually reproducing organisms, one member of each pair is derived from the maternal parent (through the ovum) and the other member is derived from the paternal parent (through the sperm). Therefore, each diploid organism contains two copies of each gene as a consequence of **biparental inheritance**, inheritance from two parents. As we shall see during our discussion of transmission genetics (Chapters 3 and 4), the members of each pair of genes, while influencing the same characteristic or trait, need not be identical. In a population of members of the same species, many different alternative forms of the same gene, called **alleles**, can exist.

TABLE 2.1 The Haploid Number of Chromosomes for a Variety of Organisms

Common Name	Scientific Name	Haploid Number
Black bread mold	*Aspergillus nidulans*	8
Broad bean	*Vicia faba*	6
Chimpanzee	*Pan troglodytes*	24
Corn	*Zea mays*	10
Cotton	*Gossypium hirsutum*	26
Dog	*Canis familiaris*	39
Fruit fly	*Drosophila melanogaster*	4
Garden pea	*Pisum sativum*	7
House mouse	*Mus musculus*	20
Human	*Homo sapiens*	23
Jimson weed	*Datura stramonium*	12
Pink bread mold	*Neurospora crassa*	7
Roundworm	*Caenorhabditis elegans*	6
Wheat	*Triticum aestivum*	21
Yeast	*Saccharomyces cerevisiae*	16
Zebrafish	*Danio rerio*	25

The concepts of haploid number, diploid number, and homologous chromosomes are important for understanding the process of meiosis. During the formation of gametes or spores, meiosis converts the diploid number of chromosomes to the haploid number. As a result, haploid gametes or spores contain precisely one member of each homologous pair of chromosomes—that is, one complete haploid set. Following fusion of two gametes at fertilization, the diploid number is reestablished; that is, the zygote contains two complete haploid sets of chromosomes. The constancy of genetic material is thus maintained from generation to generation.

There is one important exception to the concept of homologous pairs of chromosomes. In many species, one pair, consisting of the ***sex-determining chromosomes,*** is often not homologous in size, centromere placement, arm ratio, or genetic content. For example, in humans, while females carry two homologous X chromosomes, males carry one Y chromosome in addition to one X chromosome (Figure 2.4). These X and Y chromosomes are not strictly homologous. The Y is considerably smaller and lacks most of the gene loci contained on the X. Nevertheless, they contain homologous regions and behave as homologs in meiosis so that gametes produced by males receive either one X or one Y chromosome.

2.3 Mitosis Partitions Chromosomes into Dividing Cells

The process of mitosis is critical to all eukaryotic organisms. In some single-celled organisms, such as protozoans and some fungi and algae, mitosis (as a part of cell division) provides the basis for asexual reproduction. Multicellular diploid organisms begin life as single-celled fertilized eggs called **zygotes**. The mitotic activity of the zygote and the subsequent daughter cells is the foundation for the development and growth of the organism. In adult organisms, mitotic activity is the basis for wound healing and other forms of cell replacement in certain tissues. For example, the epidermal cells of the skin and the intestinal lining of humans are continuously sloughed off and replaced. Cell division also results in the continuous production of reticulocytes that eventually shed their nuclei and replenish the supply of red blood cells in vertebrates. In abnormal situations, somatic cells may lose control of cell division, and form a tumor.

The genetic material is partitioned into daughter cells during nuclear division, or **karyokinesis**. This process is quite complex and requires great precision. The chromosomes must first be exactly replicated and then accurately partitioned. The end result is the production of two daughter nuclei, each with a chromosome composition identical to that of the parent cell.

Karyokinesis is followed by cytoplasmic division, or **cytokinesis**. This less complex process requires a mechanism that partitions the volume into two parts and then encloses each new cell in a distinct plasma membrane. As the cytoplasm is reconstituted, organelles replicate themselves, arise from existing membrane structures, or are synthesized *de novo* (anew) in each cell.

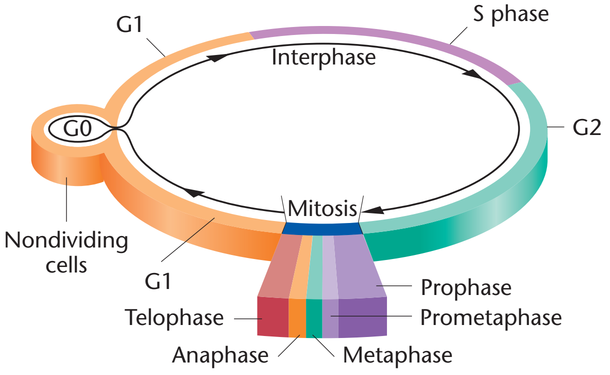


FIGURE 2.5 The stages comprising an arbitrary cell cycle. Following mitosis, cells enter the G1 stage of interphase, initiating a new cycle. Cells may become nondividing (G0) or continue through G1, where they become committed to begin DNA synthesis (S) and complete the cycle (G2 and mitosis). Following mitosis, two daughter cells are produced, and the cycle begins anew for both of them.

Following cell division, the initial size of each new daughter cell is approximately one-half the size of the parent cell. However, the nucleus of each new cell is not appreciably smaller than the nucleus of the original cell. Quantitative measurements of DNA confirm that there is an amount of genetic material in the daughter nuclei equivalent to that in the parent cell.

Interphase and the Cell Cycle

Many cells undergo a continuous alternation between division and nondivision. The events that occur from the completion of one division until the completion of the next division constitute the **cell cycle** (Figure 2.5). We will consider **interphase**, the initial stage of the cell cycle, as the interval between divisions. It was once thought that the biochemical activity during interphase was devoted solely to the cell's growth and its normal function. However, we now know that another biochemical step critical to the ensuing mitosis occurs during interphase: *the replication of the DNA of each chromosome.* This period, during which DNA is synthesized, occurs before the cell enters mitosis and is called the **S phase**. The initiation and completion of synthesis can be detected by monitoring the incorporation of radioactive precursors into DNA.

Investigations of this nature demonstrate two periods during interphase when no DNA synthesis occurs, one before and one after the S phase. These are designated **G1 (gap I)** and **G2 (gap II)**, respectively. During both of these intervals, as well as during S, intensive metabolic activity, cell growth, and cell differentiation are evident. By the end of G2, the volume of the cell has roughly doubled,

Interphase			Mitosis
G1	S	G2	M
5	7	3	1

Hours

Pro	Met	Ana	Tel
36	3	3	18

Minutes

FIGURE 2.6 The time spent in each interval of one complete cell cycle of a human cell in culture. Times vary according to cell types and conditions.

DNA has been replicated, and mitosis (M) is initiated. Following mitosis, continuously dividing cells then repeat this cycle (G1, S, G2, M) over and over, as shown in Figure 2.5.

Much is known about the cell cycle based on *in vitro* (literally, "in glass") studies. When grown in culture, many cell types in different organisms traverse the complete cycle in about 16 hours. The actual process of mitosis occupies only a small part of the overall cycle, often less than an hour. The lengths of the S and G2 phases of interphase are fairly consistent in different cell types. Most variation is seen in the length of time spent in the G1 stage. **Figure 2.6** shows the relative length of these intervals as well as the length of the stages of mitosis in a human cell in culture.

G1 is of great interest in the study of cell proliferation and its control. At a point during G1, all cells follow one of two paths. They either withdraw from the cycle, become quiescent, and enter the **G0 stage** (see Figure 2.5), or they become committed to proceed through G1, initiating DNA synthesis, and completing the cycle. Cells that enter G0 remain viable and metabolically active but are not proliferative. Cancer cells apparently avoid entering G0 or pass through it very quickly. Other cells enter G0 and never reenter the cell cycle. Still other cells in G0 can be stimulated to return to G1 and thereby reenter the cell cycle.

Cytologically, interphase is characterized by the absence of visible chromosomes. Instead, the nucleus is filled with chromatin fibers that are formed as the chromosomes uncoil and disperse after the previous mitosis [**Figure 2.7(a)**]. Once G1, S, and G2 are completed, mitosis is initiated. Mitosis is a dynamic period of vigorous and continual activity. For discussion purposes, the entire process is subdivided into discrete stages, and specific events are assigned to each one. These stages, in order of occurrence, are prophase, prometaphase, metaphase, anaphase, and telophase. They are diagrammed with corresponding photomicrographs in Figure 2.7.

Prophase

Often, over half of mitosis is spent in **prophase** [**Figure 2.7(b)**], a stage characterized by several significant occurrences. One of the early events in prophase of all animal cells is the migration of two pairs of centrioles to opposite ends of the cell. These structures are found just outside the nuclear envelope in an area of differentiated cytoplasm called the centrosome (introduced in Section 2.1). It is believed that each pair of centrioles consists of one mature unit and a smaller, newly formed daughter centriole.

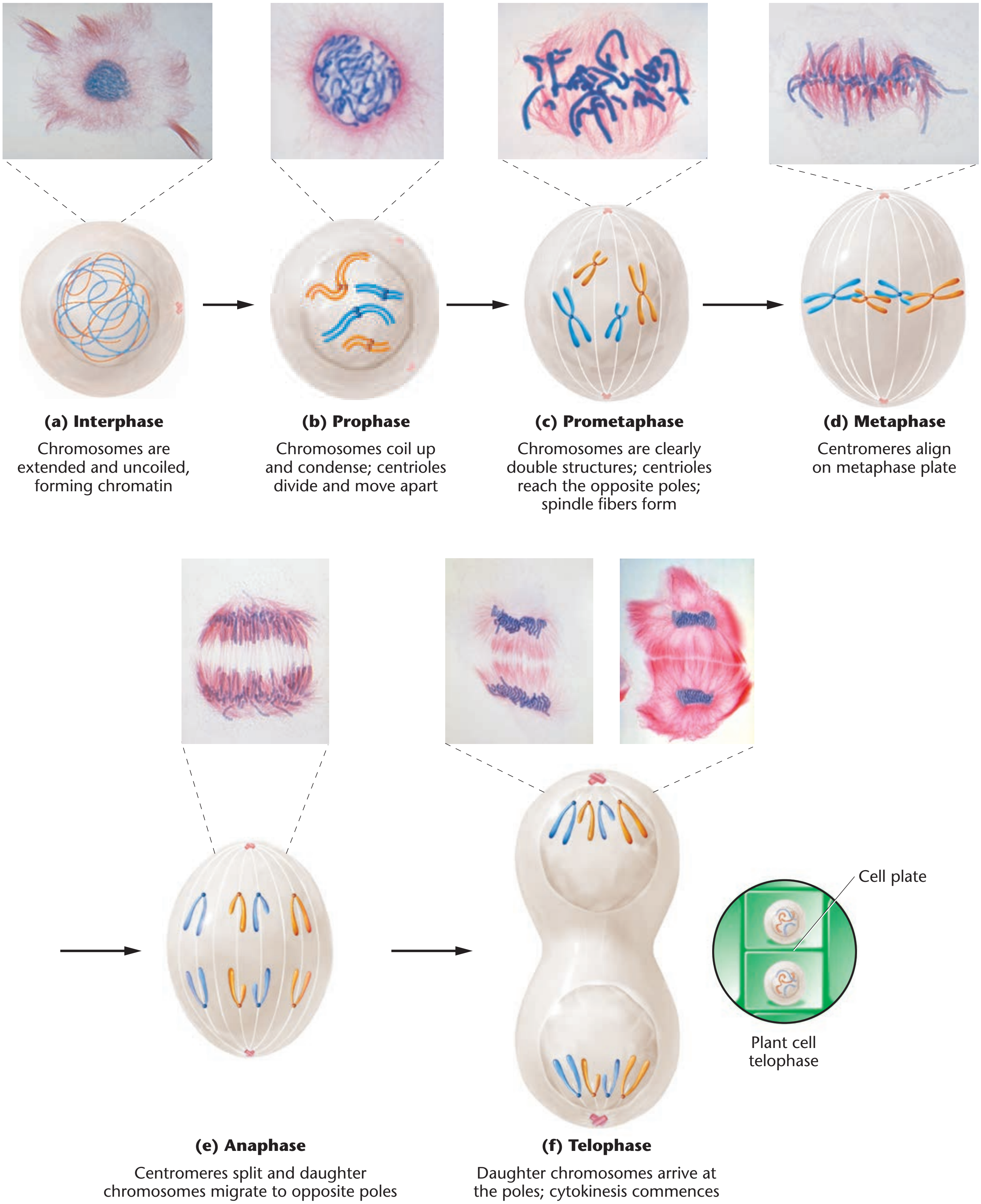
The centrioles migrate and establish poles at opposite ends of the cell. After migration, the centrosomes, in which the centrioles are localized, are responsible for organizing cytoplasmic microtubules into the spindle fibers that run between these poles, creating an axis along which chromosomal separation occurs. Interestingly, the cells of most plants (there are a few exceptions), fungi, and certain algae seem to lack centrioles. Spindle fibers are nevertheless apparent during mitosis.

As the centrioles migrate, the nuclear envelope begins to break down and gradually disappears. In a similar fashion, the nucleolus disintegrates within the nucleus. While these events are taking place, the diffuse chromatin fibers have begun to condense, until distinct thread-like structures, the chromosomes, become visible. It becomes apparent near the end of prophase that each chromosome is actually a double structure split longitudinally except at a single point of constriction, the centromere. The two parts of each chromosome are called **sister chromatids** because the DNA contained in each of them is genetically identical, having formed from a single replicative event. Sister chromatids are held together by a multi-subunit protein complex called **cohesin**. This molecular complex is originally formed between them during the S phase of the cell cycle when the DNA of each chromosome is replicated. Thus, even though we cannot see chromatids in interphase because the chromatin is uncoiled and dispersed in the nucleus, the chromosomes are already double structures, which becomes apparent in late prophase. In humans, with a diploid number of 46, a cytological preparation of late prophase reveals 46 chromosomes randomly distributed in the area formerly occupied by the nucleus.

Prometaphase and Metaphase

The distinguishing event of the two ensuing stages is the migration of every chromosome, led by its centromeric region, to the equatorial plane. The equatorial plane, also referred to as the *metaphase plate,* is the midline region of the cell, a plane that lies perpendicular to the axis established by the spindle fibers. In some descriptions, the term **prometaphase** refers to the period of chromosome movement [**Figure 2.7(c)**], and the term **metaphase** is applied strictly to the chromosome configuration following migration.

Migration is made possible by the binding of spindle fibers to the chromosome's **kinetochore**, an assembly of multilayered plates of proteins associated with the centromere. This structure forms on opposite sides of each paired centromere, in intimate association with the two sister



(a) Interphase

Chromosomes are extended and uncoiled, forming chromatin

(b) Prophase

Chromosomes coil up and condense; centrioles divide and move apart

(c) Prometaphase

Chromosomes are clearly double structures; centrioles reach the opposite poles; spindle fibers form

(d) Metaphase

Centromeres align on metaphase plate

(e) Anaphase

Centromeres split and daughter chromosomes migrate to opposite poles

(f) Telophase

Daughter chromosomes arrive at the poles; cytokinesis commences

FIGURE 2.7 Drawings depicting mitosis in an animal cell with a diploid number of 4. The events occurring in each stage are described in the text. Of the two homologous pairs of chromosomes, one pair consists of longer, metacentric members and the other of shorter, submetacentric members. The maternal chromosome and the paternal chromosome of each pair are shown in different colors. To the right of (f), a drawing of late telophase in a plant cell shows the formation of the cell plate and lack of centrioles. The cells shown in the light micrographs came from the flower of *Haemanthus,* a plant that has a diploid number of 8.

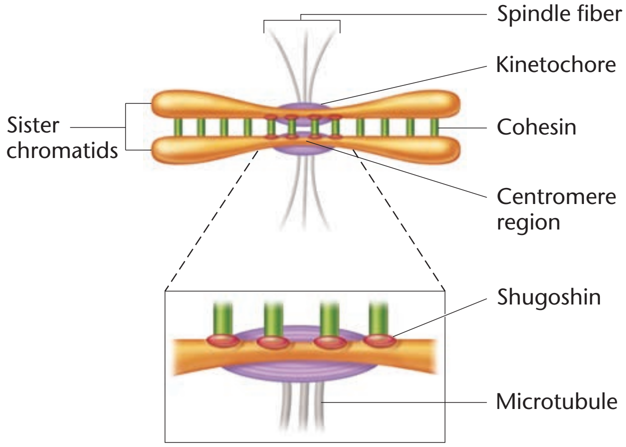


FIGURE 2.8 The depiction of the alignment, pairing, and disjunction of sister chromatids during mitosis, involving the molecular complexes cohesin and shugoshin and the enzyme separase.

chromatids. Once properly attached to the spindle fibers, cohesin is degraded by an enzyme, appropriately named *separase,* and the sister chromatid arms disjoin, except at the centromere region. A unique protein family called **shugoshin** (from the Japanese meaning "guardian spirit") protects cohesin from being degraded by separase at the centromeric regions. The involvement of the cohesin and shugoshin complexes with a pair of sister chromatids during mitosis is depicted in Figure 2.8.

We know a great deal about the molecular interactions involved in kinetechore assembly along the centromere. This is of great interest because of the consequences when mutations alter the proteins that make up the kinetechore complex. Altered kinetechore function potentially leads to errors during chromosome migration, altering the diploid content of daughter cells. A more detailed account will be presented later in the text, once we have provided more information about DNA and the proteins that make up chromatin (see Chapter 12).

We also know a great deal about spindle fibers and the mechanism responsible for their attachment to the kinetechore. Spindle fibers consist of microtubules, which themselves consist of molecular subunits of the protein tubulin. Microtubules seem to originate and "grow" out of the two centrosome regions at opposite poles of the cell. They are dynamic structures that lengthen and shorten as a result of the addition or loss of polarized tubulin subunits. The microtubules most directly responsible for chromosome migration make contact with, and adhere to, kinetochores as they grow from the centrosome region. They are referred to as ***kinetochore microtubules*** and have one end near the centrosome region (at one of the poles of the cell) and the other end anchored to the kinetochore. The number of microtubules that bind to the kinetochore varies greatly between organisms. Yeast (*Saccharomyces*) has only a single microtubule bound to each plate-like structure of the kinetochore. Mitotic cells of mammals, at the other extreme, reveal 30 to 40 microtubules bound to each portion of the kinetochore.

At the completion of metaphase, each centromere is aligned at the metaphase plate with the chromosome arms extending outward in a random array. This configuration is shown in Figure 2.7(d).

Anaphase

Events critical to chromosome distribution during mitosis occur during **anaphase**, the shortest stage of mitosis. During this phase, sister chromatids of each chromosome, held together only at their centromere regions, *disjoin* (separate) from one another—an event described as **disjunction**—and are pulled to opposite ends of the cell. For complete disjunction to occur: (1) shugoshin must be degraded, reversing its protective role; (2) the cohesin complex holding the centromere region of each sister chromosome is then cleaved by separase; and (3) sister chromatids of each chromosome are pulled toward the opposite poles of the cell (Figure 2.8). As these events proceed, each migrating chromatid is now referred to as a ***daughter chromosome***.

Movement of daughter chromosomes to the opposite poles of the cell is dependent on the kinetechore–spindle fiber attachment. Recent investigations reveal that chromosome migration results from the activity of a series of specific molecules called *motor proteins* found at several locations within the dividing cell. These proteins, described as **molecular motors**, use the energy generated by the hydrolysis of ATP. Their effect on the activity of microtubules serves ultimately to shorten the spindle fibers, drawing the chromosomes to opposite ends of the cell. The centromeres of each chromosome *appear* to lead the way during migration, with the chromosome arms trailing behind. Several models have been proposed to account for the shortening of spindle fibers. They share in common the selective removal of tubulin subunits at the ends of the spindle fibers. The removal process is accomplished by the molecular motor proteins described above.

The location of the centromere determines the shape of the chromosome during separation, as you saw in Figure 2.3. The steps that occur during anaphase are critical in providing each subsequent daughter cell with an identical set of chromosomes. In human cells, there would now be 46 chromosomes at each pole, one from each original sister pair. Figure 2.7(e) shows anaphase prior to its completion.

Telophase

Telophase is the final stage of mitosis and is depicted in Figure 2.7(f). At its beginning, two complete sets of chromosomes are present, one set at each pole. The most significant event of this stage is cytokinesis, the division or partitioning of the cytoplasm. Cytokinesis is essential if two new cells are to be produced from one cell. The mechanism of cytokinesis differs greatly in plant and animal cells, but the end result is the same: two new cells are produced. In plant cells, a ***cell plate*** is synthesized and laid down across the region of the metaphase plate. Animal cells, however, undergo a constriction of the cytoplasm, much as a loop of string might be tightened around the middle of a balloon.

It is not surprising that the process of cytokinesis varies in different organisms. Plant cells, which are more regularly shaped and structurally rigid, require a mechanism for depositing new cell wall material around the plasma membrane. The cell plate laid down during telophase becomes a structure called the ***middle lamella***. Subsequently, the primary and secondary layers of the cell wall are deposited between the cell membrane and middle lamella in each of the resulting daughter cells. In animals, complete constriction of the cell membrane produces the ***cell furrow*** characteristic of newly divided cells.

Other events necessary for the transition from mitosis to interphase are initiated during late telophase. They generally constitute a reversal of events that occurred during prophase. In each new cell, the chromosomes begin to uncoil and become diffuse chromatin once again, while the nuclear envelope reforms around them, the spindle fibers disappear, and the nucleolus gradually reforms and becomes visible in the nucleus during early interphase. At the completion of telophase, the cell enters interphase.

Cell-Cycle Regulation and Checkpoints

The cell cycle, culminating in mitosis, is fundamentally the same in all eukaryotic organisms. This similarity in many diverse organisms suggests that the cell cycle is governed by a genetically regulated program that has been conserved throughout evolution. Because disruption of this regulation may underlie the uncontrolled cell division characterizing malignancy, interest in how genes regulate the cell cycle is particularly strong.

A mammoth research effort over the past 20 years has paid high dividends, and we now have knowledge of many genes involved in the control of the cell cycle. This work was recognized by the awarding of the 2001 Nobel Prize in Medicine or Physiology to Lee Hartwell, Paul Nurse, and Tim Hunt. As with other studies of genetic control over essential biological processes, investigation has focused on the discovery of mutations that interrupt the cell cycle and on the effects of those mutations. As we shall return to this subject in much greater detail later in the text during our consideration of the molecular basis of cancer (see Chapter 24), what follows is a very brief overview.

Many mutations are now known that exert an effect at one or another stage of the cell cycle. First discovered in yeast, but now evident in all organisms, including humans, such mutations were originally designated as ***cell division cycle (cdc) mutations***. The normal products of many of the mutated genes are enzymes called **kinases** that can add phosphates to other proteins. They serve as "master control" molecules functioning in conjunction with proteins called **cyclins**. Cyclins bind to these kinases (creating *cyclin-dependent kinases*), activating them at appropriate times during the cell cycle. Activated kinases then phosphorylate other target proteins that regulate the progress of the cell cycle. The study of *cdc* mutations has established that the cell cycle contains at least three **cell-cycle checkpoints,** where the processes culminating in normal mitosis are monitored, or "checked," by these master control molecules before the next stage of the cycle is allowed to commence.

The importance of cell-cycle control and these checkpoints can be demonstrated by considering what happens when this regulatory system is impaired. Let's assume, for example, that the DNA of a cell has incurred damage leading to one or more mutations impairing cell-cycle control. If allowed to proceed through the cell cycle, this genetically altered cell would divide uncontrollably—a key step in the development of a cancer cell. If, instead, the cell cycle is arrested at one of the checkpoints, the cell can repair the DNA damage or permanently stop the cell from dividing, thereby preventing its potential malignancy. The specific checkpoints will be discussed in more detail later in the text (Chapter 24, Cancer Genetics).

NOW SOLVE THIS

2.1 With the initial appearance of the feature we call "Now Solve This," a short introduction is in order. The feature occurs several times in this and all ensuing chapters, each time providing a problem related to the discussion just presented. A "Hint" is then offered that may help you solve the problem. Here is the first problem:

(a) If an organism has a diploid number of 16, how many chromatids are visible at the end of mitotic prophase?

(b) How many chromosomes are moving to each pole during anaphase of mitosis?

■ **HINT:** *This problem involves an understanding of what happens to each pair of homologous chromosomes during mitosis, asking you to apply your understanding of chromosome behavior to an organism with a diploid number of 16. The key to its solution is your awareness that throughout mitosis, the members of each homologous pair do not pair up, but instead behave independently.*

2.4 Meiosis Creates Haploid Gametes and Spores and Enhances Genetic Variation in Species

Whereas in diploid organisms, mitosis produces two daughter cells with full diploid complements, **meiosis** produces gametes or spores that are characterized by only one haploid set of chromosomes. During sexual reproduction, haploid gametes then combine at fertilization to reconstitute the

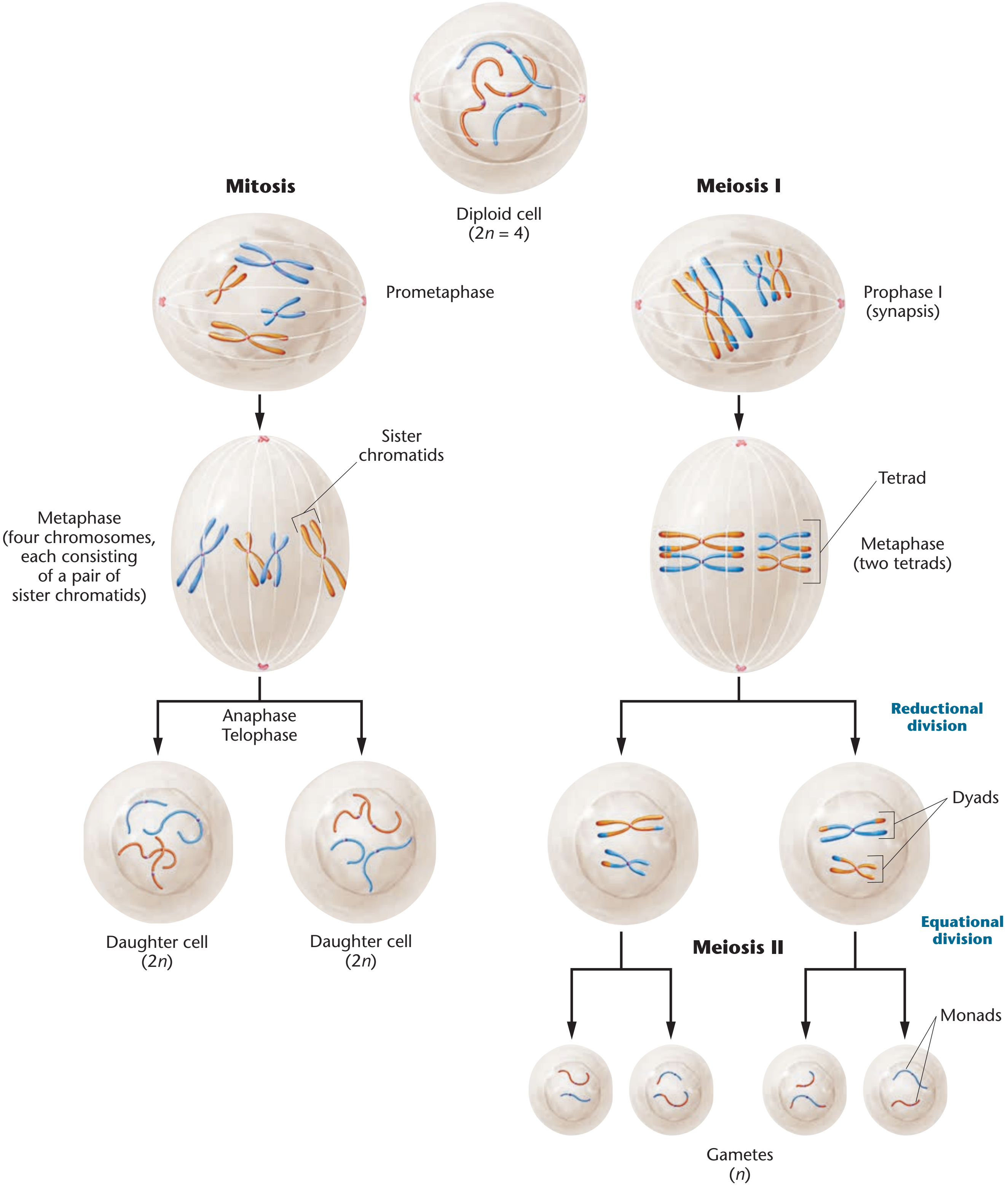


FIGURE 2.9 Overview of the major events and outcomes of mitosis and meiosis. As in Figure 2.7, two pairs of homologous chromosomes are followed.

diploid complement found in parental cells. Figure 2.9 compares the two processes by following two pairs of homologous chromosomes. Meiosis must be highly specific since, by definition, haploid gametes or spores must contain precisely one member of each homologous pair of chromosomes. When successfully completed, meiosis provides the basis for maintaining genetic continuity from generation to generation.

Another major accomplishment of meiosis is to ensure that during sexual reproduction an enormous amount of genetic variation is produced among members of a species.

Such variation occurs in two forms. First, meiosis produces gametes with many unique combinations of maternally and paternally derived chromosomes among the haploid complement, thus ensuring that following fertilization, a large number of unique chromosome combinations are possible. As we will see (Chapter 3), this process is the underlying basis of Mendel's principles of segregation and independent assortment. The second source of variation is created by the meiotic event referred to as **crossing over**, which results in genetic exchange between members of each homologous pair of chromosomes prior to one or the other finding its way into a haploid gamete or spore. This creates intact chromosomes that are mosaics of the maternal and paternal homologs from which they arise, further enhancing genetic variation. Sexual reproduction therefore significantly reshuffles the genetic material, producing highly diverse offspring.

NOW SOLVE THIS

2.2 An organism has a diploid number of 16 in a primary oocyte. (a) How many tetrads are present in the first meiotic prophase? (b) How many dyads are present in the second meiotic prophase? (c) How many monads migrate to each pole during the second meiotic anaphase?

- **HINT:** *This problem involves an understanding of what happens to the maternal and paternal members of each pair of homologous chromosomes during meiosis, asking you to apply your understanding of chromosome behavior in an organism with a diploid number of 16. The key to its solution is your awareness that maternal and paternal homologs synapse during meiosis. Once each chromatid has duplicated, creating a tetrad in the early phases of meiosis, each original pair behaves as a unit and leads to two dyads during anaphase I.*

For more practice, see Problems 25–30.

Meiosis: Prophase I

As in mitosis, the process in meiosis begins with a diploid cell duplicating its genetic material in the interphase stage preceding chromosome division. To achieve haploidy, two divisions are thus required. The meiotic achievements, as described above, are largely dependent on the behavior of chromosomes during the initial stage of the first division, called ***prophase I.*** Recall that in mitosis the paternally and maternally derived members of each homologous pair of chromosomes behave autonomously during division. Each chromosome is duplicated, creating genetically identical sister chromatids, and subsequently, one chromatid of each pair is distributed to each new cell. The major difference in meiosis is that once the chromatin characterizing interphase has condensed into visible structures, the homologous chromosomes are not autonomous but are instead seen to be paired up, having undergone the process called **synapsis**. Figure 2.10 illustrates this process as well as the ensuing events of prophase I. Each synapsed pair of homologs is initially called a **bivalent,** and the number of bivalents is equal to the haploid number. In Figure 2.10, we have depicted two homologous pairs of chromosomes and thus two bivalents. As the homologs condense and shorten, each bivalent gives rise to a unit called a **tetrad,** consisting of two pairs of sister chromatids, each of which is joined at a common centromere. Remember that one pair of sister chromatids is maternally derived and the other pair is paternally derived. The presence of tetrads is visible evidence that *both* homologs have, in fact, duplicated. As prophase I progresses, each pair of sister chromatids within a tetrad is seen to pull apart. However, one or more areas remain in contact where chromatids are intertwined. Each such area, called a **chiasma** (pl., chiasmata), is thought to represent a point where **nonsister chromatids** (one paternal and one maternal chromatid) have undergone genetic exchange through the process of crossing over. Since crossing over is thought to occur one or more times in each tetrad, mosaic chromosomes are routinely created during every meiotic event. During the final period of prophase I, the nucleolus and nuclear envelope break down, and the two centromeres of each tetrad attach to the recently formed spindle fibers.

Metaphase, Anaphase, and Telophase I

The remainder of the meiotic process is depicted in Figure 2.11. After meiotic prophase I, stages similar to those of mitosis occur. In the first division, ***metaphase I,*** the chromosomes have maximally shortened and thickened.

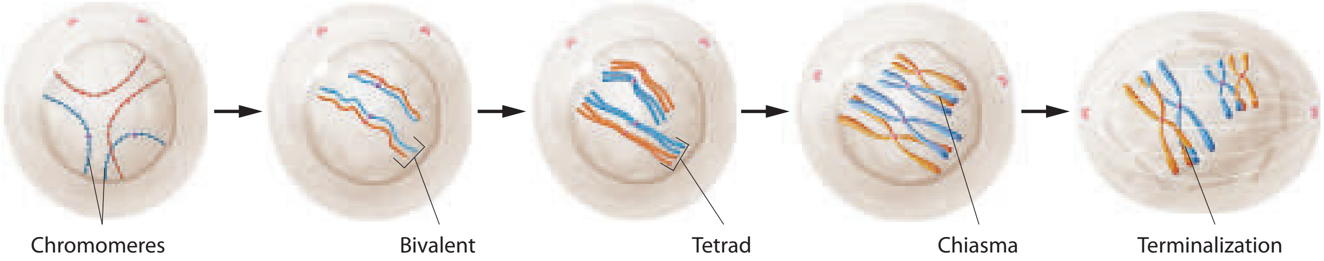


FIGURE 2.10 The changes in chromosome structures during prophase I, which characterize each of the events of the process.

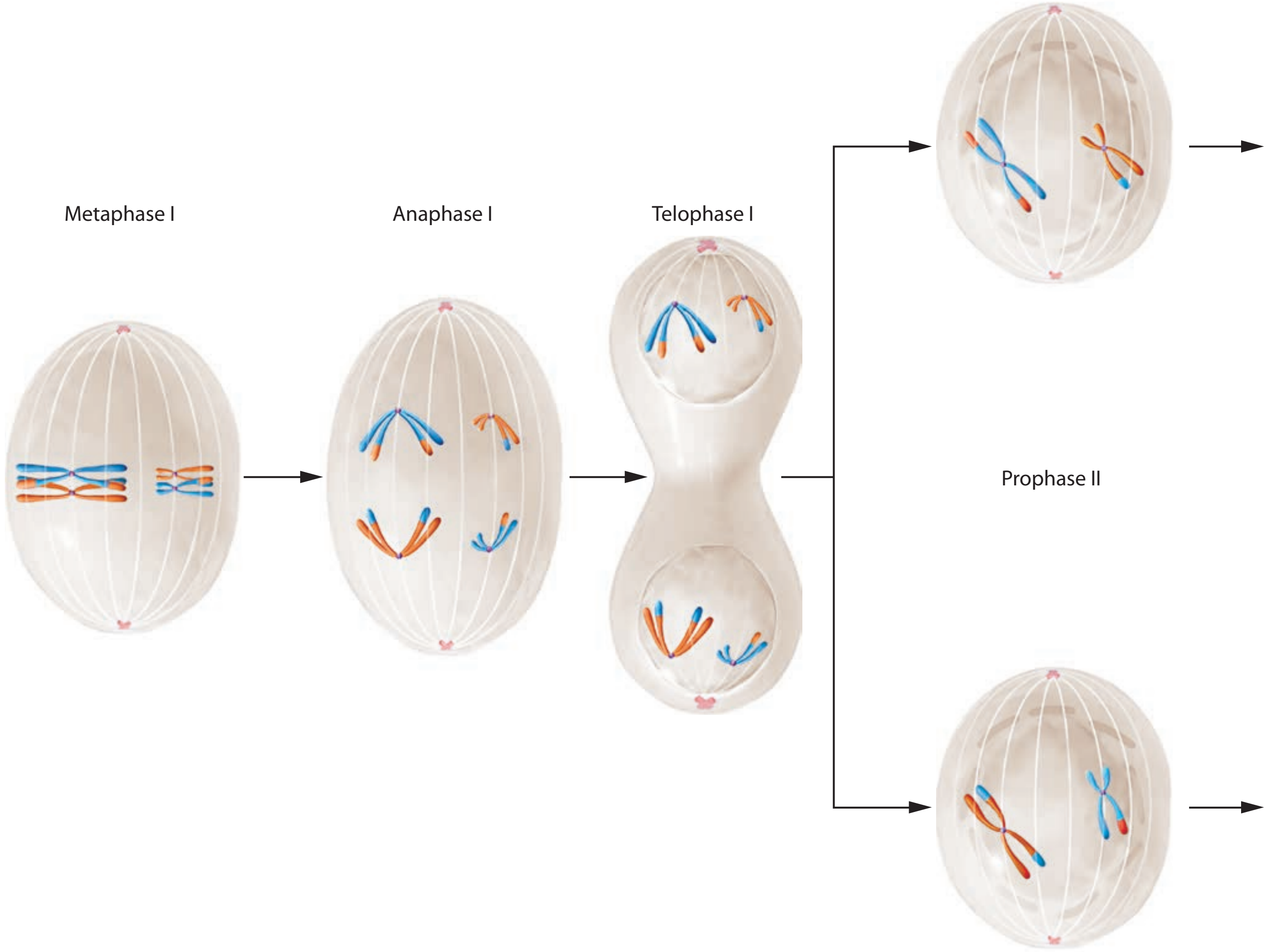


FIGURE 2.11 The major events in meiosis in an animal cell with a diploid number of 4, beginning with metaphase I. Note that the combination of chromosomes in the cells produced following telophase II is dependent on the random alignment of each tetrad and dyad on the equatorial plate during metaphase I and metaphase II. Several other combinations, which are not shown, can also be formed. The events depicted here are described in the text.

The terminal chiasmata of each tetrad are visible and appear to be the major factor holding the nonsister chromatids together. Each tetrad interacts with spindle fibers, facilitating its movement to the metaphase plate. The alignment of each tetrad prior to the first anaphase is random: Half of the tetrad (one of the dyads) will subsequently be pulled by spindle fibers to one or the other pole, and the other half will be pulled to the opposite pole.

During the stages of meiosis I, a single centromeric region holds each pair of sister chromatids together. It appears as a single unit, and a kinetechore forms around each one. As in our discussion of mitosis (see Figure 2.8), cohesin plays the major role in keeping sister chromatids together. At ***anaphase I,*** cohesin is degraded between sister chromatids, except at the centromere region, which, as in mitosis, is protected by a shugoshin complex. Then, one-half of each tetrad (a **dyad**) is pulled toward each pole of the dividing cell. Because this process effectively reduces the number of centromeres by half, it is referred to as a *reductional division*. This separation process is the physical basis of disjunction, the separation of homologous chromosomes from one another. Occasionally, errors in meiosis occur and separation is not achieved. The term **nondisjunction** describes such an error. At the completion of the normal anaphase I, a series of dyads equal to the haploid number is present at each pole.

If crossing over had not occurred in the first meiotic prophase, each dyad at each pole would consist solely of either paternal or maternal chromatids. However, the exchanges produced by crossing over create mosaic chromatids of paternal and maternal origin.

In many organisms, ***telophase I*** reveals a nuclear membrane forming around the dyads. In this case, the nucleus next

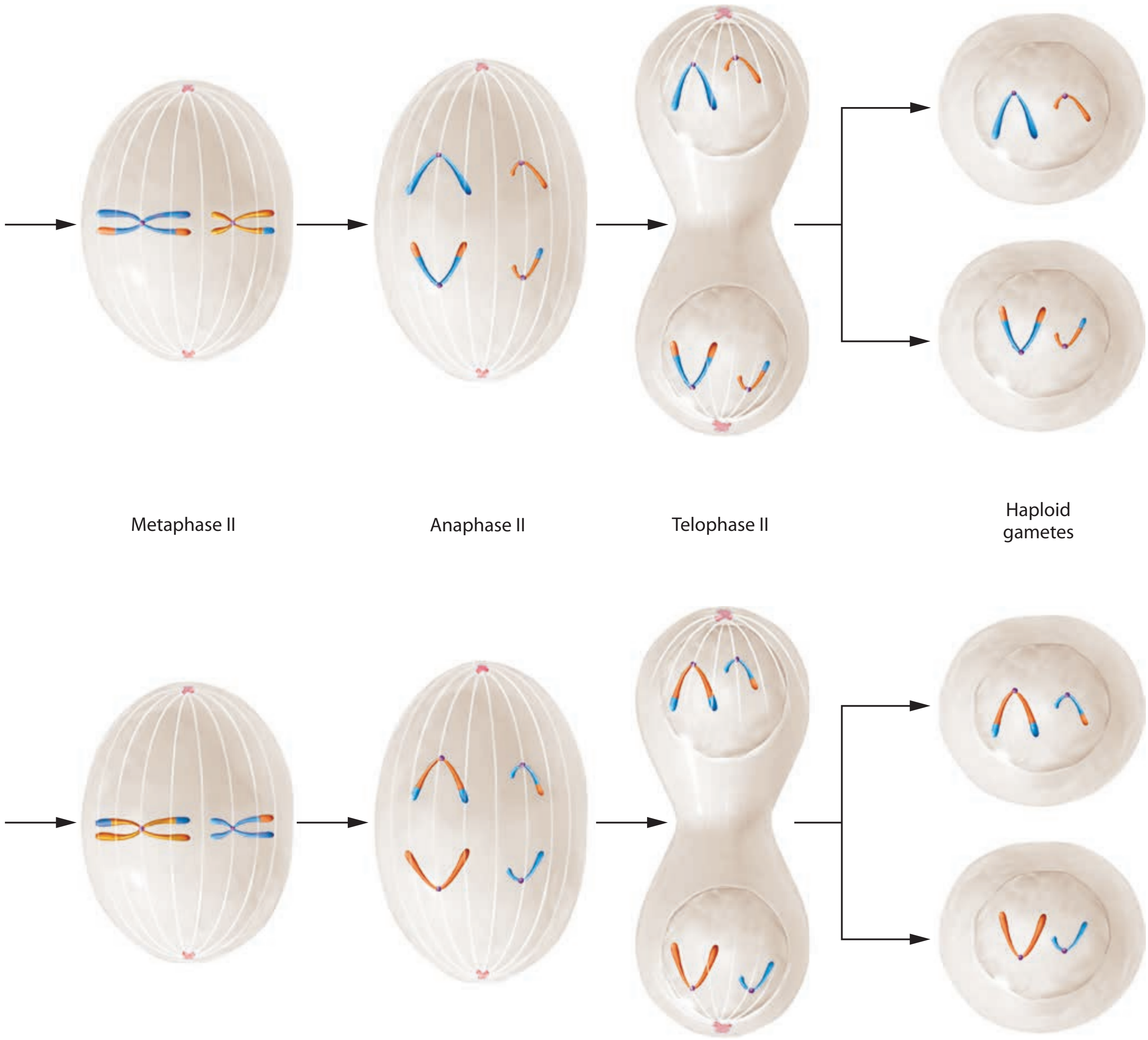


FIGURE 2.11 *(Continued)*

enters into a short interphase period. If interphase occurs, the chromosomes do not replicate because they already consist of two chromatids. In other organisms, the cells go directly from anaphase I to meiosis II. In general, meiotic telophase is much shorter than the corresponding stage in mitosis.

The Second Meiotic Division

A second division, referred to as ***meiosis II,*** is essential if each gamete or spore is to receive only one chromatid from each original tetrad. The stages characterizing meiosis II are shown on the right side of Figure 2.11. During ***prophase II,*** each dyad is composed of one pair of sister chromatids attached by the common centromeric region. During ***metaphase II,*** the centromeres are positioned on the equatorial plate. When the shugoshin complex is degraded, the centromeres separate, ***anaphase II*** is initiated, and the sister chromatids of each dyad are pulled to opposite poles. Because the number of dyads is equal to the haploid number, ***telophase II*** reveals one member of each pair of homologous chromosomes present at each pole. Each chromosome is now a monad. Because the number of centromeres is not reduced in number in the two resulting cells, the process is referred to as an *equational division*.

Following cytokinesis in telophase II, four haploid gametes may result from a single meiotic event. At the conclusion of meiosis II, not only has the haploid state been achieved, but if crossing over has occurred, each monad contains a combination of maternal and paternal genetic information. As a result, the offspring produced by any gamete will receive a mixture of genetic information originally present in his or her grandparents. Meiosis thus significantly increases the level of genetic variation in each ensuing generation.

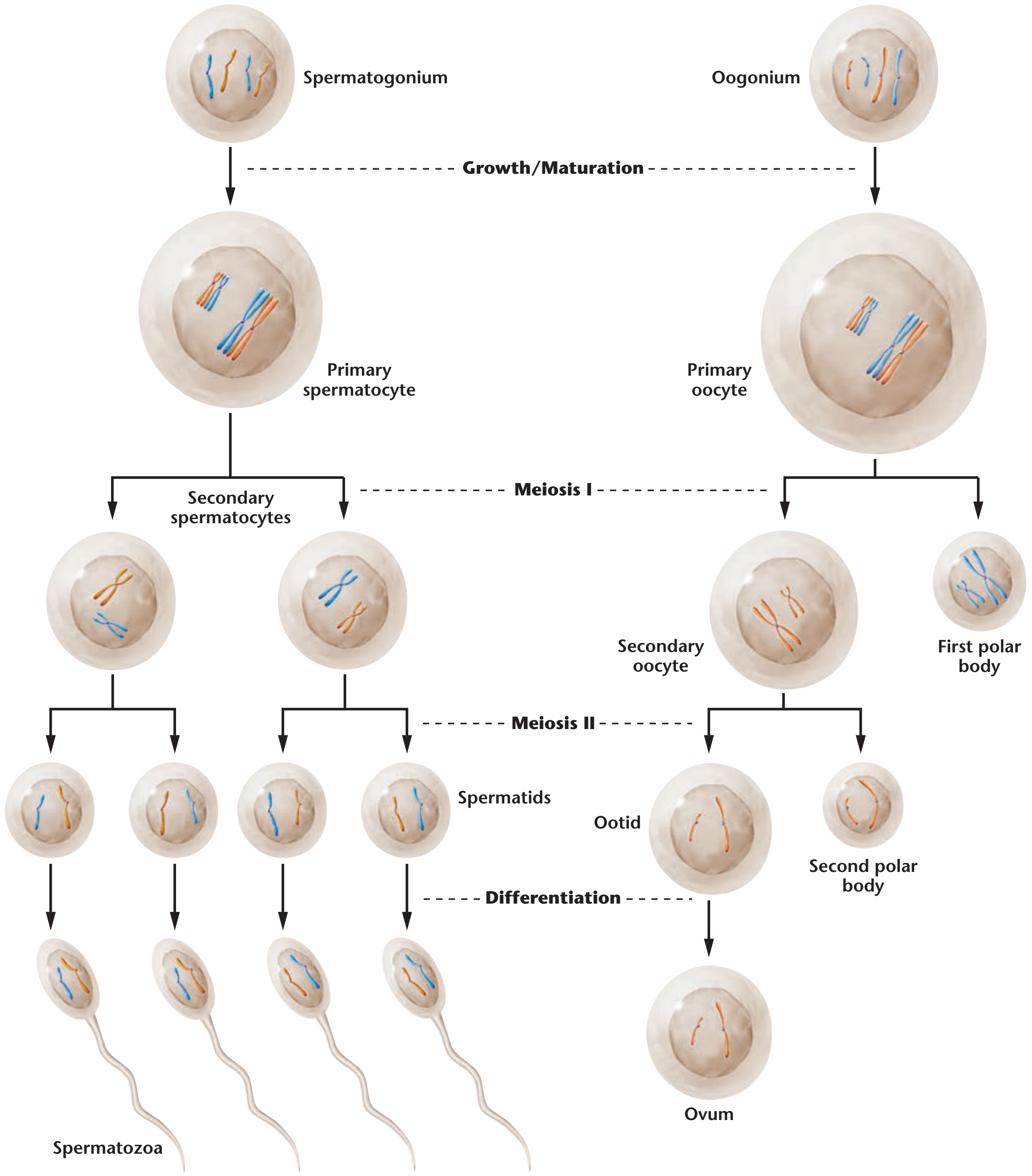


FIGURE 2.12 Spermatogenesis and oogenesis in animal cells.

2.5 The Development of Gametes Varies in Spermatogenesis Compared to Oogenesis

Although events that occur during the meiotic divisions are similar in all cells participating in gametogenesis in most animal species, there are certain differences between the production of a male gamete (spermatogenesis) and a female gamete (oogenesis). Figure 2.12 summarizes these processes.

Spermatogenesis takes place in the testes, the male reproductive organs. The process begins with the enlargement of an undifferentiated diploid germ cell called a ***spermatogonium***. This cell grows to become a ***primary spermatocyte,*** which undergoes the first meiotic division. The products of this division, called ***secondary spermatocytes,*** contain a haploid number of dyads. The secondary spermatocytes then undergo meiosis II, and each of these cells produces two haploid ***spermatids***. Spermatids go through a series of developmental changes, ***spermiogenesis,*** to become highly specialized, motile ***spermatozoa,*** or ***sperm***. All sperm cells produced during spermatogenesis contain the haploid number of chromosomes and equal amounts of cytoplasm.

Spermatogenesis may be continuous or may occur periodically in mature male animals; its onset is determined by the species' reproductive cycles. Animals that reproduce year-round produce sperm continuously, whereas those whose breeding period is confined to a particular season produce sperm only during that time.

In animal ***oogenesis,*** the formation of ***ova*** (sing. **ovum**), or eggs, occurs in the ovaries, the female reproductive organs. The daughter cells resulting from the two meiotic divisions of this process receive equal amounts of genetic material, but they do *not* receive equal amounts of cytoplasm. Instead, during each division, almost all the cytoplasm of the ***primary oocyte,*** itself derived from the ***oogonium,*** is concentrated in one of the two daughter cells. The concentration of cytoplasm is necessary because a major function of the mature ovum is to nourish the developing embryo following fertilization.

During anaphase I in oogenesis, the tetrads of the primary oocyte separate, and the dyads move toward opposite poles. During telophase I, the dyads at one pole are pinched off with very little surrounding cytoplasm to form the ***first polar body***. The first polar body may or may not divide again to produce two small haploid cells. The other daughter cell produced by this first meiotic division contains most of the cytoplasm and is called the ***secondary oocyte***. The mature ovum will be produced from the secondary oocyte during the second meiotic division. During this division, the cytoplasm of the secondary oocyte again divides unequally, producing an **ootid** and a **second polar body**. The ootid then differentiates into the mature ovum.

Unlike the divisions of spermatogenesis, the two meiotic divisions of oogenesis may not be continuous. In some animal species, the second division may directly follow the first. In others, including humans, the first division of all oocytes begins in the embryonic ovary but arrests in prophase I. Many years later, meiosis resumes in each oocyte just prior to its ovulation. The second division is completed only after fertilization.

NOW SOLVE THIS

2.3 Examine Figure 2.12, which shows oogenesis in animal cells. Will the genotype of the second polar body (derived from meiosis II) always be identical to that of the ootid? Why or why not?

■ **HINT:** *This problem involves an understanding of meiosis during oogenesis, asking you to demonstrate your knowledge of polar bodies. The key to its solution is to take into account that crossing over occurred between each pair of homologs during meiosis I.*

2.6 Meiosis Is Critical to Sexual Reproduction in All Diploid Organisms

The process of meiosis is critical to the successful sexual reproduction of all diploid organisms. It is the mechanism by which the diploid amount of genetic information is reduced to the haploid amount. In animals, meiosis leads to the formation of gametes, whereas in plants haploid spores are produced, which in turn lead to the formation of haploid gametes.

Each diploid organism stores its genetic information in the form of homologous pairs of chromosomes. Each pair consists of one member derived from the maternal parent and one from the paternal parent. Following meiosis, haploid cells potentially contain either the paternal or the maternal representative of every homologous pair of chromosomes. However, the process of crossing over, which occurs in the first meiotic prophase, further reshuffles the alleles between the maternal and paternal members of each homologous pair, which then segregate and assort independently into gametes. These events result in the great amount of genetic variation present in gametes.

It is important to touch briefly on the significant role that meiosis plays in the life cycles of fungi and plants. In many fungi, the predominant stage of the life cycle consists of haploid vegetative cells. They arise through meiosis and proliferate by mitotic cell division. In multicellular plants, the life cycle alternates between the diploid **sporophyte stage**

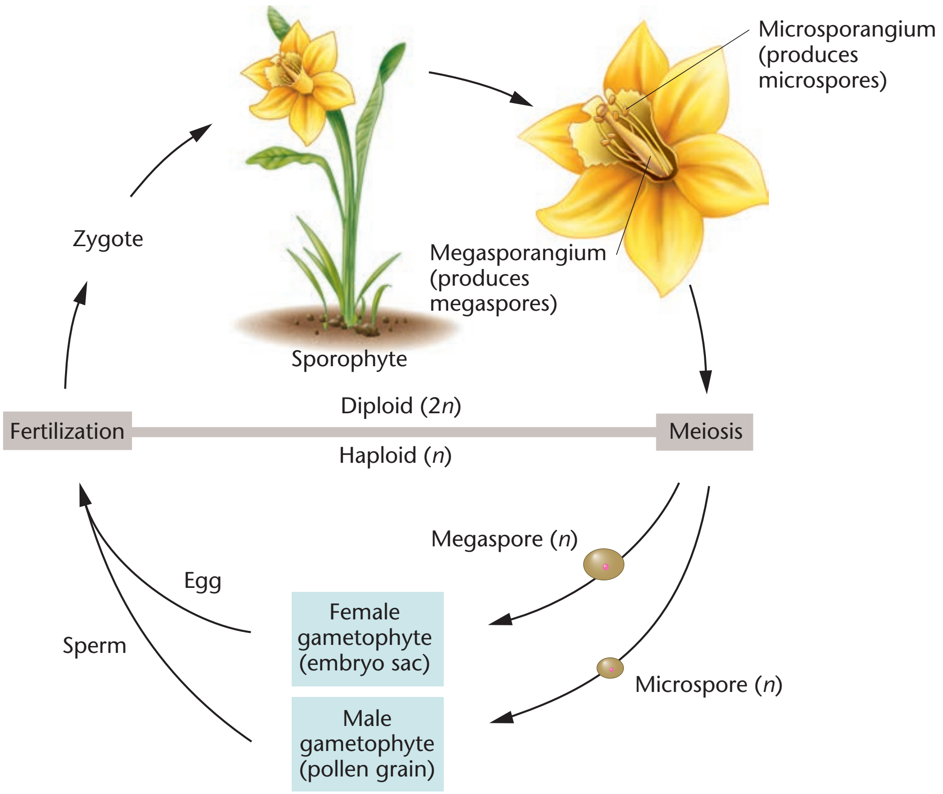

Microsporangium (produces microspores)
Zygote
Megasporangium (produces megaspores)
Sporophyte
Diploid (2*n*)
Fertilization
Meiosis
Haploid (*n*)
Megaspore (*n*)
Egg
Female gametophyte (embryo sac)
Sperm
Microspore (*n*)
Male gametophyte (pollen grain)

FIGURE 2.13 Alternation of generations between the diploid sporophyte (2*n*) and the haploid gametophyte (*n*) in a multicellular plant. The processes of meiosis and fertilization bridge the two phases of the life cycle. In angiosperms (flowering plants), like the one shown here, the sporophyte stage is the predominant phase.

and the haploid ***gametophyte stage*** (Figure 2.13). While one or the other predominates in different plant groups during this "alternation of generations," the processes of meiosis and fertilization constitute the "bridges" between the sporophyte and gametophyte stages. Therefore, meiosis is an essential component of the life cycle of plants.

2.7 Electron Microscopy Has Revealed the Physical Structure of Mitotic and Meiotic Chromosomes

Thus far in this chapter, we have focused on mitotic and meiotic chromosomes, emphasizing their behavior during cell division and gamete formation. An interesting question is why chromosomes are invisible during interphase but visible during the various stages of mitosis and meiosis. Studies using electron microscopy clearly show why this is the case.

Recall that, during interphase, only dispersed chromatin fibers are present in the nucleus [Figure 2.14(a)]. Once mitosis begins, however, the fibers coil and fold, condensing into typical mitotic chromosomes [Figure 2.14(b)]. If the fibers comprising a mitotic chromosome are loosened, the areas of greatest spreading reveal individual fibers similar to those seen in interphase chromatin [Figure 2.14(c)]. Very few fiber ends seem to be present, and in some cases, none can be seen. Instead, individual fibers always seem to loop back into the interior. Such fibers are obviously twisted and coiled around one another, forming the regular pattern of folding in the mitotic chromosome. Starting in late telophase of mitosis and continuing during G1 of interphase, chromosomes unwind to form the long fibers characteristic of chromatin, which consist of DNA and associated proteins, particularly proteins called *histones*. It is in this physical arrangement that DNA can most efficiently function during transcription and replication.

Electron microscopic observations of metaphase chromosomes in varying degrees of coiling led Ernest DuPraw to postulate the **folded-fiber model**, shown in Figure 2.14(c). During metaphase, each chromosome consists of two sister chromatids joined at the centromeric region. Each arm of the chromatid appears to be a single fiber wound much like a skein of yarn. The fiber is composed of tightly coiled double-stranded DNA and protein. An orderly coiling–twisting–condensing

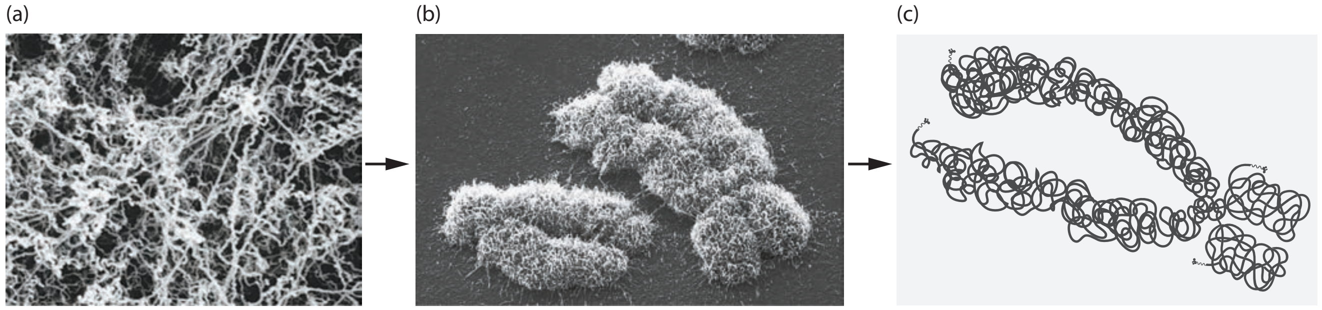
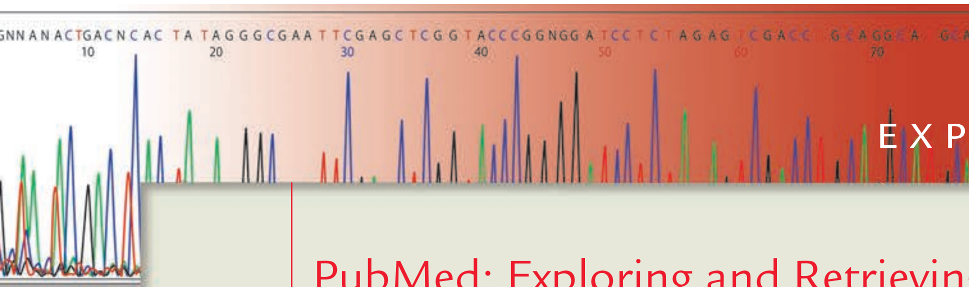


FIGURE 2.14 Comparison of (a) the chromatin fibers characteristic of the interphase nucleus with (b) metaphase chromosomes that are derived from chromatin during mitosis. Part (c) diagrams a mitotic chromosome, showing how chromatin is condensed to produce it. Part (a) is a transmission electron micrograph and part (b) is a scanning electron micrograph.

process appears to facilitate the transition of the interphase chromatin into the more condensed mitotic chromosomes. Geneticists believe that during the transition from interphase to prophase, a 5000-fold compaction occurs in the length of DNA within the chromatin fiber! This process must be extremely precise given the highly ordered and consistent appearance of mitotic chromosomes in all eukaryotes. Note particularly in the micrographs the clear distinction between the sister chromatids constituting each chromosome. They are joined only by the common centromere that they share prior to anaphase. We will return to this general topic later in the text when we consider chromosome structure in further detail (see Chapter 12).



EXPLORING GENOMICS

PubMed: Exploring and Retrieving Biomedical Literature

Mastering Genetics Visit the Study Area: Exploring Genomics

PubMed is an Internet-based search system developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. Using PubMed, one can access over 26 million citations for publications in over 5600 biomedical journals. The full text of many of the articles can be obtained electronically through college or university libraries, and some journals (such as *Proceedings of the National Academy of Sciences USA; Genome Biology;* and *Science*) provide free public access to articles within certain time frames.

In this exercise, we will explore PubMed to answer questions about relationships between tubulin, human cancers, and cancer therapies.

Exercise I – Tubulin, Cancer, and Mitosis

In this chapter we were introduced to tubulin and the dynamic behavior of microtubules during the cell cycle. Cancer cells are characterized by continuous and uncontrolled mitotic divisions.

Is it possible that tubulin and microtubules contribute to the development of cancer? Could these important structures be targets for cancer therapies?

1. To begin your search for the answers, access the PubMed site at http://www.ncbi.nlm.nih.gov/pubmed/.
2. In the search box, type “tubulin cancer” and then click the “Search” button to perform the search.
3. Select several research papers and read the abstracts.

To answer the question about tubulin’s association with cancer, you may want to limit your search to fewer papers, perhaps those that are review articles. To do this, click the “Review” link under the Article Types category on the left side of the page.

Explore some of the articles, as abstracts or as full text, available in your library or by free public access. Prepare a brief report or verbally share your experiences with your class. Describe two of the most important things you learned during your exploration, and identify the information sources you encountered during the search.

CASE STUDY Timing is everything

Over a period of two years, a man in his early 20s received a series of intermittent chemotherapy and radiotherapy treatments for Hodgkin disease. During this therapy, he and his wife were unable to initiate a pregnancy. The man had a series of his semen samples examined at a fertility clinic. The findings revealed that shortly after each treatment very few mature sperm were present, and abnormal chromosome numbers were often observed in developing spermatocytes. However, such chromosome abnormalities disappeared about 40 days after treatment, and normal sperm reappeared about 74 days post-treatment.

1. How might a genetic counselor explain the time-related differences in sperm production and the appearance and subsequent disappearance of chromosomal abnormalities?
2. Do you think that exposure to chemotherapy and radiotherapy would cause more problems to spermatocytes than to mature sperm?
3. Prior to treatment, should the physician(s) involved have been ethically obligated to recommend genetic counseling? What advice regarding fertility might have been suggested?

For further reading, see: Harel, S., et al., 2011. Management of fertility in patients treated for Hodgkin's lymphoma. *Haematologica* 2011. 96: 1692–99.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. The structure of cells is elaborate and complex, with most components involved directly or indirectly with genetic processes.
2. In diploid organisms, chromosomes exist in homologous pairs, where each member is identical in size, centromere placement, and gene loci. One member of each pair is derived from the maternal parent, and the other from the paternal parent.
3. Mitosis and meiosis are mechanisms by which cells distribute the genetic information contained in their chromosomes to progeny cells in a precise, orderly fashion.
4. Mitosis, which is but one part of the cell cycle, is subdivided into discrete stages that initially depict the condensation of chromatin into the diploid number of chromosomes. Each chromosome first appears as a double structure, consisting of a pair of identical sister chromatids joined at a common centromere. As mitosis proceeds, centromeres split and sister chromatids are pulled apart by spindle fibers and directed toward opposite poles of the cell. Cytoplasmic division then occurs, creating two new cells with the identical genetic information contained in the progenitor cell.
5. Meiosis converts a diploid cell into haploid gametes or spores, making sexual reproduction possible. As a result of chromosome duplication, two subsequent meiotic divisions are required to achieve haploidy, whereby each haploid cell receives one member of each homologous pair of chromosomes.
6. There is a major difference between meiosis in males and in females. Spermatogenesis partitions the cytoplasmic volume equally and produces four haploid sperm cells. Oogenesis, on the other hand, collects the bulk of cytoplasm in one egg cell and reduces the other haploid products to polar bodies. The extra cytoplasm in the egg contributes to zygote development following fertilization.
7. Meiosis results in extensive genetic variation by virtue of the exchange of chromosome segments during crossing over between maternal and paternal chromatids and by virtue of the random separation of maternal and paternal chromatids into gametes. In addition, meiosis plays an important role in the life cycles of fungi and plants, serving as the bridge between alternating generations.
8. Mitotic chromosomes are produced as a result of the coiling and condensation of chromatin fibers of interphase into the characteristic form of chromatids.

INSIGHTS AND SOLUTIONS

This appearance of "Insights and Solutions" begins a feature that will have great value to you as a student. From this point on, "Insights and Solutions" precedes the "Problems and Discussion Questions" at each chapter's end to provide sample problems and solutions that demonstrate approaches you will find useful in genetic analysis. The insights you gain by working through the sample problems will improve your ability to solve the ensuing problems in each chapter.

1. In an organism with a diploid number of $2n = 6$, how many individual chromosomal structures will align on the metaphase plate during (a) mitosis, (b) meiosis I, and (c) meiosis II? Describe each configuration.

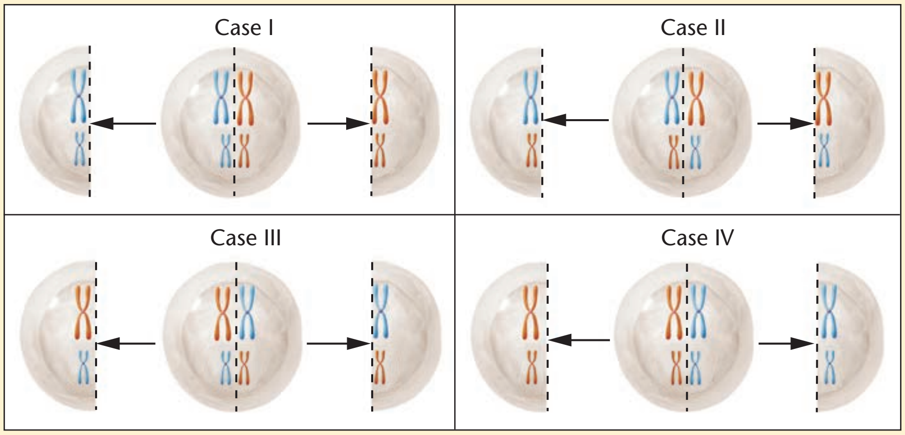
Solution: (a) Remember that in mitosis, homologous chromosomes do not synapse, so there will be six double structures, each consisting of a pair of sister chromatids. In other words, the number of structures is equivalent to the diploid number.

(b) In meiosis I, the homologs have synapsed, reducing the number of structures to three. Each is called a tetrad and consists of two pairs of sister chromatids.

(c) In meiosis II, the same number of structures exist (three), but in this case they are called dyads. Each dyad is a pair of sister chromatids. When crossing over has occurred, each chromatid may contain parts of one of its nonsister chromatids, obtained during exchange in prophase I.

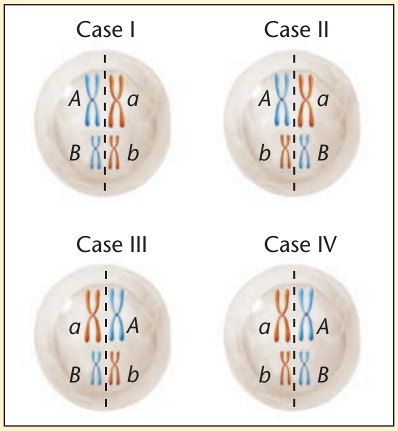
2. Disregarding crossing over, draw all possible alignment configurations that can occur during metaphase for the chromosomes shown in Figure 2.11.

Solution: As shown in the diagram below, four configurations are possible when $n = 2$.



3. For the chromosomes in Problem 2, assume that each of the larger chromosomes has a different allele for a given gene, *A* OR *a*, as shown. Also assume that each of the smaller chromosomes has a different allele for a second gene, *B* OR *b*. Calculate the probability of generating each possible combination of these alleles (*AB*, *Ab*, *aB*, *ab*) following meiosis I.

Solution: As shown in the accompanying diagram:



Case I	*AB* and *ab*	**Total:**	$AB = 2(p = 1/4)$
Case II	*Ab* and *aB*		$Ab = 2(p = 1/4)$
Case III	*aB* and *Ab*		$aB = 2(p = 1/4)$
Case IV	*ab* and *AB*		$ab = 2(p = 1/4)$

4. How many different chromosome configurations can occur following meiosis I if three different pairs of chromosomes are present ($n = 3$)?

Solution: If $n = 3$, then eight different configurations would be possible. The formula 2^n, where n equals the haploid number, represents the number of potential alignment patterns. As we will see in (Chapter 3), these patterns are produced according to the Mendelian postulate of *segregation,* and they serve as the physical basis of another Mendelian postulate called *independent assortment.*

5. Describe the composition of a meiotic tetrad during prophase I, assuming no crossover event has occurred. What impact would a single crossover event have on this structure?

Solution: Such a tetrad contains four chromatids, existing as two pairs. Members of each pair are sister chromatids. They are held together by a common centromere. Members of one pair are maternally derived, whereas members of the other are paternally derived. Maternal and paternal members are called nonsister chromatids. A single crossover event has the effect of exchanging a portion of a maternal and a paternal chromatid, leading to a chiasma, where the two involved chromatids overlap physically in the tetrad. The process of exchange is referred to as crossing over.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on how chromosomes are distributed during cell division, both in dividing somatic cells (mitosis) and in gamete- and spore-forming cells (meiosis). We found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, answer the following questions.
(a) How do we know that chromosomes exist in homologous pairs?
(b) How do we know that DNA replication occurs during interphase, not early in mitosis?
(c) How do we know that mitotic chromosomes are derived from chromatin?
2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 14. All of these pertain to conceptual issues involving mitosis or meiosis. Based on these concepts, write a short essay that contrasts mitosis and meiosis, including their respective roles in organisms, the mechanisms by which they achieve their respective outcomes, and the consequences should either process fail to be executed with absolute fidelity.
3. What role do the following cellular components play in the storage, expression, or transmission of genetic information: (a) chromatin, (b) nucleolus, (c) ribosome, (d) mitochondrion, (e) centriole, (f) centromere?
4. Discuss the concepts of homologous chromosomes, diploidy, and haploidy. What characteristics do two homologous chromosomes share?
5. If two chromosomes of a species are the same length and have similar centromere placements and yet are not homologous, what is different about them?
6. Describe the events that characterize each stage of mitosis.
7. How are chromosomes named on the basis of their centromere placement?
8. Contrast telophase in plant and animal mitosis.
9. Describe the phases of the cell cycle and the events that characterize each phase.
10. Define and discuss these terms: (a) synapsis, (b) bivalents, (c) chiasmata, (d) crossing over, (e) chromomeres, (f) sister chromatids, (g) tetrads, (h) dyads, (i) monads.
11. Contrast the genetic content and the origin of sister versus nonsister chromatids during their earliest appearance in prophase I of meiosis. How might the genetic content of these change by the time tetrads have aligned at the equatorial plate during metaphase I?
12. Given the end results of the two types of division, why is it necessary for homologs to pair during meiosis and not desirable for them to pair during mitosis?
13. Contrast spermatogenesis and oogenesis. What is the significance of the formation of polar bodies?
14. Explain why meiosis leads to significant genetic variation while mitosis does not.
15. A diploid cell contains three pairs of homologous chromosomes designated C1 and C2, M1 and M2, and S1 and S2. No crossing over occurs. What combinations of chromosomes are possible in (a) daughter cells following mitosis, (b) cells undergoing the first meiotic metaphase, (c) haploid cells following both divisions of meiosis?
16. Considering Problem 15, predict the number of different haploid cells that could be produced by meiosis if a fourth chromosome pair (W1 and W2) were added.
17. During oogenesis in an animal species with a haploid number of 6, one dyad undergoes nondisjunction during meiosis II. Following the second meiotic division, this dyad ends up intact in the ovum. How many chromosomes are present in (a) the mature ovum and (b) the second polar body? (c) Following fertilization by a normal sperm, what chromosome condition is created?
18. What is the probability that, in an organism with a haploid Number of 10, a sperm will be formed that contains all 10 chromosomes whose centromeres were derived from maternal homologs?
19. The nuclear DNA content of a single sperm cell in *Drosophila melanogaster* is approximately 0.18 picogram. What would be the expected nuclear DNA content of a primary spermatocyte in *Drosophila*? What would be the expected nuclear DNA content of a somatic cell (non-sex cell) in the G1 phase? What would be the expected nuclear DNA content of a somatic cell at metaphase?
20. Describe the role of meiosis in the life cycle of a vascular plant.
21. Contrast the chromatin fiber with the mitotic chromosome. How are the two structures related?
22. Describe the "folded-fiber" model of the mitotic chromosome.
23. You are given a metaphase chromosome preparation (a slide) from an unknown organism that contains 12 chromosomes. Two that are clearly smaller than the rest appear identical in length and centromere placement. Describe all that you can about these chromosomes.
24. If one follows 50 primary oocytes in an animal through their various stages of oogenesis, how many secondary oocytes would be formed? How many first polar bodies would be formed? How many ootids would be formed? If one follows 50 primary spermatocytes in an animal through their various stages of spermatogenesis, how many secondary spermatocytes would be formed? How many spermatids would be formed?

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

As part of the "Problems and Discussion Questions" section in this and each subsequent chapter, we shall present a number of "Extra-Spicy" genetics problems. We have chosen to set these apart in order to identify problems that are particularly challenging. You may be asked to examine and assess actual data, to design genetics experiments, or to engage in cooperative learning. Like genetic varieties of peppers, some of these experiences are just spicy and some are very hot. We hope that you will enjoy the challenges that they pose.

For Problems **25–30**, consider a diploid cell that contains three pairs of chromosomes designated AA, BB, and CC. Each pair contains a maternal and a paternal member (e.g., A^m and A^p). Using these designations, demonstrate your understanding of mitosis and meiosis by drawing chromatid combinations as requested. Be sure to indicate when chromatids are paired as a result of replication and/or synapsis. You may wish to use a large piece of brown manila wrapping paper or a cut-up paper grocery bag for this project and to work in partnership with another student. We recommend cooperative learning as an efficacious way to develop the skills you will need for solving the problems presented throughout this text.

25. In mitosis, what chromatid combination(s) will be present during metaphase? What combination(s) will be present at each pole at the completion of anaphase?

26. During meiosis I, assuming no crossing over, what chromatid combination(s) will be present at the completion of prophase I? Draw all possible alignments of chromatids as migration begins during early anaphase.

27. Are there any possible combinations present during prophase of meiosis II other than those that you drew in Problem 26? If so, draw them.

28. Draw all possible combinations of chromatids during the early phases of anaphase in meiosis II.

29. Assume that during meiosis I none of the *C* chromosomes disjoin at metaphase, but they separate into dyads (instead of monads) during meiosis II. How would this change the alignments that you constructed during the anaphase stages in meiosis I and II? Draw them.

30. Assume that each gamete resulting from Problem 29 fuses, in fertilization, with a normal haploid gamete. What combinations will result? What percentage of zygotes will be diploid, containing one paternal and one maternal member of each chromosome pair?

31. A species of cereal rye (*Secale cereale*) has a chromosome number of 14, while a species of Canadian wild rye (*Elymus canadensis*) has a chromosome number of 28. Sterile hybrids can be produced by crossing *Secale* with *Elymus*.
(a) What would be the expected chromosome number in the somatic cells of the hybrids?
(b) Given that none of the chromosomes pair at meiosis I in the sterile hybrid (Hang and Franckowlak, 1984), speculate on the anaphase I separation patterns of these chromosomes.

32. An interesting procedure has been applied for assessing the chromosomal balance of potential secondary oocytes for use in human *in vitro* fertilization. Using fluorescence *in situ* hybridization (FISH), Kuliev and Verlinsky (2004) were able to identify individual chromosomes in first polar bodies and thereby infer the chromosomal makeup of “sister” oocytes. Assume that when examining a first polar body you saw that it had one copy (dyad) of each chromosome but two dyads of chromosome 21. What would you expect to be the chromosomal 21 complement in the secondary oocyte? What consequences are likely in the resulting zygote, if the secondary oocyte was fertilized?

33. Assume that you were examining a first polar body and noted that it had one copy (dyad) of each chromosome except chromosome 21. Chromosome 21 was completely absent. What would you expect to be the chromosome 21 complement (only with respect to chromosome 21) in the secondary oocyte? What consequences are likely in the resulting zygote if the secondary oocyte was fertilized?

34. Kuliev and Verlinsky (2004) state that there was a relatively high number of separation errors at meiosis I. In these cases the centromere underwent a premature division, occurring at meiosis I rather than meiosis II. Regarding chromosome 21, what would you expect to be the chromosome 21 complement in the secondary oocyte in which you saw a single chromatid (monad) for chromosome 21 in the first polar body? If this secondary oocyte was involved in fertilization, what would be the expected consequences?

3

Mendelian Genetics

Gregor Johann Mendel, who in 1866 put forward the major postulates of transmission genetics as a result of experiments with the garden pea.

CHAPTER CONCEPTS

- Inheritance is governed by information stored in discrete unit factors called genes.
- Genes are transmitted from generation to generation on vehicles called chromosomes.
- Chromosomes, which exist in pairs in diploid organisms, provide the basis of biparental inheritance.
- During gamete formation, chromosomes are distributed according to postulates first described by Gregor Mendel, based on his nineteenth-century research with the garden pea.
- Mendelian postulates prescribe that homologous chromosomes segregate from one another and assort independently with other segregating homologs during gamete formation.
- Genetic ratios, expressed as probabilities, are subject to chance deviation and may be evaluated statistically.
- The analysis of pedigrees allows predictions concerning the genetic nature of human traits.

Although inheritance of biological traits has been recognized for thousands of years, the first significant insights into how it takes place only occurred about 150 years ago. In 1866, Gregor Johann Mendel published the results of a series of experiments that would lay the foundation for the formal discipline of genetics. Mendel's work went largely unnoticed until the turn of the twentieth century, but eventually, the concept of the gene as a distinct hereditary unit was established. Since then, the ways in which genes, as segments of chromosomes, are transmitted to offspring and control traits have been clarified. Research continued unabated throughout the twentieth century and into the present—indeed, studies in genetics, most recently at the molecular level, have remained at the forefront of biological research since the early 1900s.

When Mendel began his studies of inheritance using *Pisum sativum,* the garden pea, chromosomes and the role and mechanism of meiosis were totally unknown. Nevertheless, he determined that discrete units of inheritance exist and predicted their behavior in the formation of gametes. Subsequent investigators, with access to cytological data, were able to relate their own observations of chromosome behavior during meiosis and Mendel's principles of inheritance. Once this correlation was recognized, Mendel's postulates were accepted as the basis for the study of what is known as **transmission genetics**—how genes are transmitted from parents to offspring. These principles were derived directly from Mendel's experimentation.

3.1 Mendel Used a Model Experimental Approach to Study Patterns of Inheritance

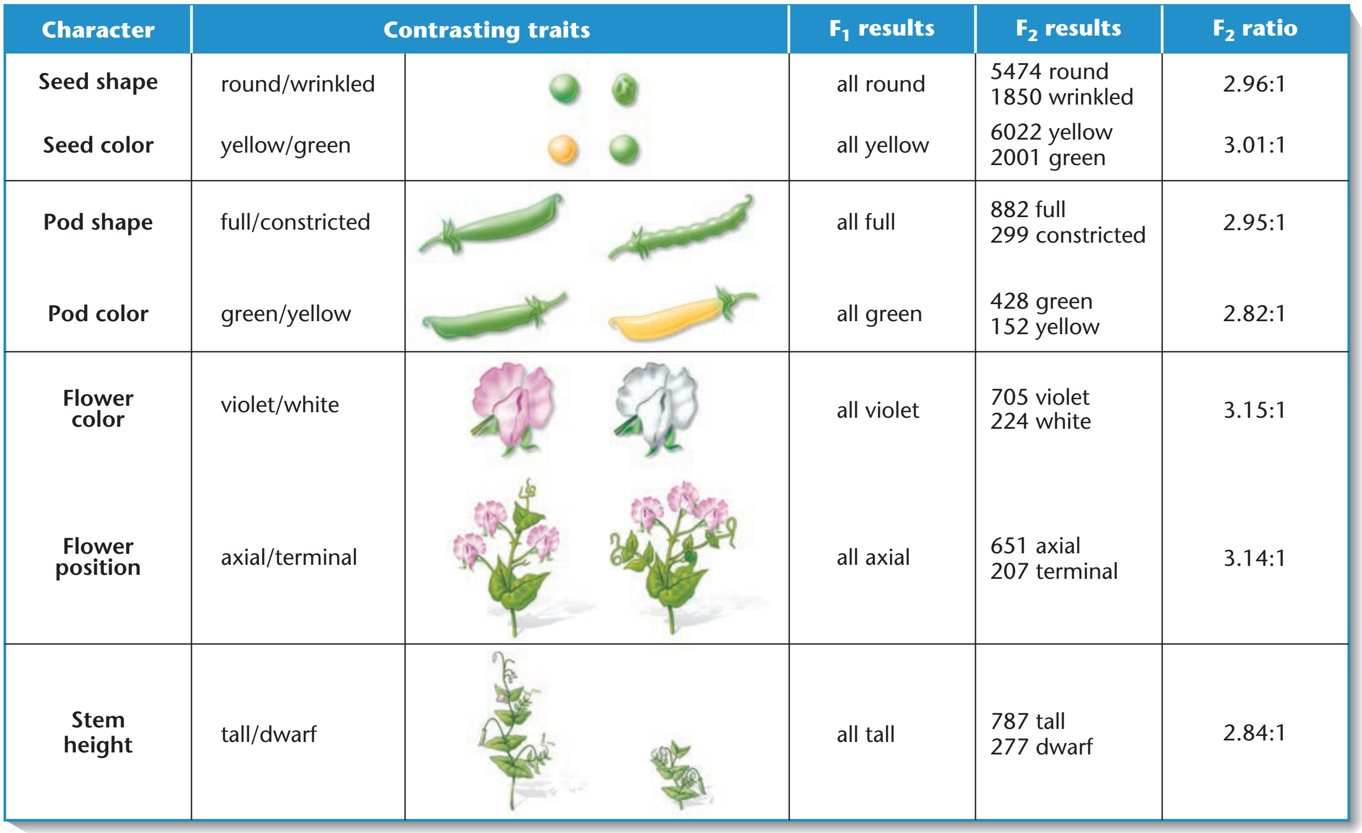
Johann Mendel was born in 1822 to a peasant family in the Central European village of Heinzendorf. An excellent student in high school, he studied philosophy for several years afterward and in 1843, taking the name Gregor, was admitted to the Augustinian Monastery of St. Thomas in Brno, now part of the Czech Republic. In 1849, he was relieved of pastoral duties, and from 1851 to 1853, he attended the University of Vienna, where he studied physics and botany. He returned to Brno in 1854, where he taught physics and natural science for the next 16 years. Mendel received support from the monastery for his studies and research throughout his life.

In 1856, Mendel performed his first set of hybridization experiments with the garden pea, launching the research phase of his career. His experiments continued until 1868, when he was elected abbot of the monastery. Although he retained his interest in genetics, his new responsibilities demanded most of his time. In 1884, Mendel died of a kidney disorder. The local newspaper paid him the following tribute:

> His death deprives the poor of a benefactor, and mankind at large of a man of the noblest character, one who was a warm friend, a promoter of the natural sciences, and an exemplary priest.

Mendel first reported the results of some simple genetic crosses between certain strains of the garden pea in 1865. Although his was not the first attempt to provide experimental evidence pertaining to inheritance, Mendel's success where others had failed can be attributed, at least in part, to his elegant experimental design and analysis.

Mendel showed remarkable insight into the methodology necessary for good experimental biology. First, he chose an organism that was easy to grow and to hybridize artificially. The pea plant is self-fertilizing in nature, but it is easy to cross-breed experimentally. It reproduces well and grows to maturity in a single season. Mendel followed seven visible features (we refer to them as characters, or characteristics), each represented by two contrasting forms, or **traits** (**Figure 3.1**). For the character stem height, for example, he experimented with the traits *tall* and *dwarf.* He selected



Character	Contrasting traits		F_1 results	F_2 results	F_2 ratio
Seed shape	round/wrinkled		all round	5474 round 1850 wrinkled	2.96:1
Seed color	yellow/green		all yellow	6022 yellow 2001 green	3.01:1
Pod shape	full/constricted		all full	882 full 299 constricted	2.95:1
Pod color	green/yellow		all green	428 green 152 yellow	2.82:1
Flower color	violet/white		all violet	705 violet 224 white	3.15:1
Flower position	axial/terminal		all axial	651 axial 207 terminal	3.14:1
Stem height	tall/dwarf		all tall	787 tall 277 dwarf	2.84:1

FIGURE 3.1 Seven pairs of contrasting traits and the results of Mendel's seven monohybrid crosses of the garden pea (*Pisum sativum*). In each case, pollen derived from plants exhibiting one trait was used to fertilize the ova of plants exhibiting the other trait. In the F_1 generation, one of the two traits was exhibited by all plants. The contrasting trait reappeared in approximately 1/4 of the F_2 plants.

six other contrasting pairs of traits involving seed shape and color, pod shape and color, and flower color and position. From local seed merchants, Mendel obtained true-breeding strains, those in which each trait appeared unchanged generation after generation in self-fertilizing plants.

There were several other reasons for Mendel's success. In addition to his choice of a suitable organism, he restricted his examination to one or very few pairs of contrasting traits in each experiment. He also kept accurate quantitative records, a necessity in genetic experiments. From the analysis of his data, Mendel derived certain postulates that have become the principles of transmission genetics.

The results of Mendel's experiments went unappreciated until the turn of the century, well after his death. However, once Mendel's publications were rediscovered by geneticists investigating the function and behavior of chromosomes, the implications of his postulates were immediately apparent. He had discovered the basis for the transmission of hereditary traits!

3.2 The Monohybrid Cross Reveals How One Trait Is Transmitted from Generation to Generation

Mendel's simplest crosses involved only one pair of contrasting traits. Each such experiment is called a **monohybrid cross**. A monohybrid cross is made by mating true-breeding individuals from two parent strains, each exhibiting one of the two contrasting forms of the character under study. Initially, we examine the first generation of offspring of such a cross, and then we consider the offspring of **selfing**, that is, of self-fertilization of individuals from this first generation. The original parents constitute the **P_1**, or **parental generation**; their offspring are the F_1, or **first filial generation**; the individuals resulting from the selfed F_1 generation are the **F_2**, or **second filial generation**; and so on.

The cross between true-breeding pea plants with tall stems and dwarf stems is representative of Mendel's monohybrid crosses. *Tall* and *dwarf* are contrasting traits of the character of stem height. Unless tall or dwarf plants are crossed together or with another strain, they will undergo self-fertilization and breed true, producing their respective traits generation after generation. However, when Mendel crossed tall plants with dwarf plants, the resulting F_1 generation consisted of only tall plants. When members of the F_1 generation were selfed, Mendel observed that 787 of 1064 F_2 plants were tall, while 277 of 1064 were dwarf. Note that in this cross (Figure 3.1), the dwarf trait disappeared in the F_1 generation, only to reappear in the F_2 generation.

Genetic data are usually expressed and analyzed as ratios. In this particular example, many identical P_1 crosses were made and many F_1 plants—all tall—were produced. As noted, of the 1064 F_2 offspring, 787 were tall and 277 were dwarf—a ratio of approximately 2.8:1.0, or about 3:1.

Mendel made similar crosses between pea plants exhibiting each of the other pairs of contrasting traits; the results of these crosses are shown in Figure 3.1. In every case, the outcome was similar to the tall/dwarf cross just described. For the character of interest, all F_1 offspring expressed the same trait exhibited by one of the parents, but in the F_2 offspring, an approximate ratio of 3:1 was obtained. That is, three-fourths looked like the F_1 plants, while one-fourth exhibited the contrasting trait, which had disappeared in the F_1 generation.

We note one further aspect of Mendel's monohybrid crosses. In each cross, the F_1 and F_2 patterns of inheritance were similar regardless of which P_1 plant served as the source of pollen (sperm) and which served as the source of the ovum (egg). The crosses could be made either way—pollination of dwarf plants by tall plants, or vice versa. Crosses made in both these ways are called **reciprocal crosses**. Therefore, the results of Mendel's monohybrid crosses were not sex dependent.

To explain these results, Mendel proposed the existence of particulate *unit factors* for each trait. He suggested that these factors serve as the basic units of heredity and are passed unchanged from generation to generation, determining various traits expressed by each individual plant. Using these general ideas, Mendel proceeded to hypothesize precisely how such factors could account for the results of the monohybrid crosses.

Mendel's First Three Postulates

Using the consistent pattern of results in the monohybrid crosses, Mendel derived the following three postulates, or principles, of inheritance.

1. UNIT FACTORS IN PAIRS
   Genetic characters are controlled by unit factors existing in pairs in individual organisms.
   In the monohybrid cross involving tall and dwarf stems, a specific **unit factor** exists for each trait. Each diploid individual receives one factor from each parent. Because the factors occur in pairs, three combinations are possible: two factors for tall stems, two factors for dwarf stems, or one of each factor. Every individual possesses one of these three combinations, which determines stem height.

2. DOMINANCE/RECESSIVENESS
   When two unlike unit factors responsible for a single character are present in a single individual, one unit factor is dominant to the other, which is said to be recessive.
   In each monohybrid cross, the trait expressed in the F_1 generation is controlled by the dominant unit factor. The trait not expressed is controlled by the recessive unit factor. The terms dominant and recessive are also

used to designate traits. In this case, tall stems are said to be dominant over recessive dwarf stems.

3. SEGREGATION

During the formation of gametes, the paired unit factors separate, or segregate, randomly so that each gamete receives one or the other with equal likelihood.

If an individual contains a pair of like unit factors (e.g., both specific for tall), then all its gametes receive one of that same kind of unit factor (in this case, tall). If an individual contains unlike unit factors (e.g., one for tall and one for dwarf), then each gamete has a 50 percent probability of receiving either the tall or the dwarf unit factor.

These postulates provide a suitable explanation for the results of the monohybrid crosses. Let's use the tall/dwarf cross to illustrate. Mendel reasoned that P_1 tall plants contained identical paired unit factors, as did the P_1 dwarf plants. The gametes of tall plants all receive one tall unit factor as a result of segregation. Similarly, the gametes of dwarf plants all receive one dwarf unit factor. Following fertilization, all F_1 plants receive one unit factor from each parent—a tall factor from one and a dwarf factor from the other—reestablishing the paired relationship, but because tall is dominant to dwarf, all F_1 plants are tall.

When F_1 plants form gametes, the postulate of segregation demands that each gamete randomly receives either the tall *or* dwarf unit factor. Following random fertilization events during F_1 selfing, four F_2 combinations will result with equal frequency:

1. tall/tall
2. tall/dwarf
3. dwarf/tall
4. dwarf/dwarf

Combinations (1) and (4) will clearly result in tall and dwarf plants, respectively. According to the postulate of dominance/recessiveness, combinations (2) and (3) will both yield tall plants. Therefore, the F_2 is predicted to consist of 3/4 tall and 1/4 dwarf, or a ratio of 3:1. This is approximately what Mendel observed in his cross between tall and dwarf plants. A similar pattern was observed in each of the other monohybrid crosses (Figure 3.1).

Modern Genetic Terminology

To analyze the monohybrid cross and Mendel's first three postulates, we must first introduce several new terms as well as a symbol convention for the unit factors. Traits such as tall or dwarf are physical expressions of the information contained in unit factors. The physical expression of a trait is the **phenotype** of the individual. Mendel's unit factors represent units of inheritance called **genes** by modern geneticists. For any given character, such as plant height, the phenotype is determined by alternative forms of a single gene, called **alleles**. For example, the unit factors representing tall and dwarf are alleles determining the height of the pea plant.

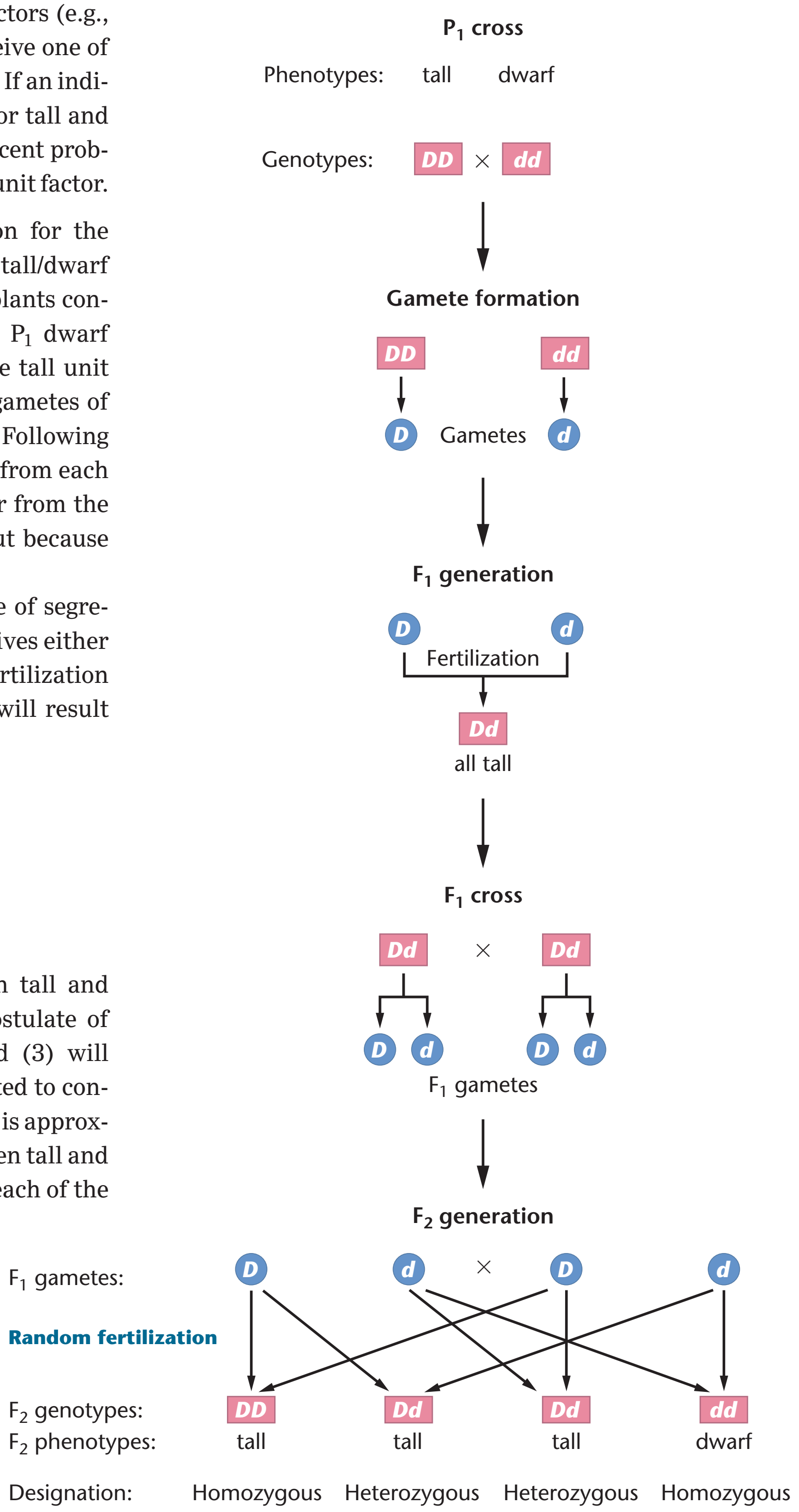


FIGURE 3.2 The monohybrid cross between tall (*D*) and dwarf (*d*) pea plants. Individuals are shown in rectangles, and gametes are shown in circles.

Geneticists have several different systems for using symbols to represent genes. Later in the text (see Chapter 4), we will review a number of these conventions, but for now, we will adopt one to use consistently throughout this chapter. According to this convention, the first letter of the recessive trait symbolizes the character in question; in lowercase italic, it designates the allele for the recessive trait, and in uppercase italic, it designates the allele for the dominant trait. Thus for Mendel's pea plants, we use *d* for the *d*warf allele and *D* for the tall allele. When alleles are written in pairs to represent the two unit factors present in any individual (*DD*, *Dd*, or *dd*), the resulting symbol is called the **genotype**. The genotype designates the genetic makeup of an individual for the trait or traits it describes, whether the individual is haploid or diploid. By reading the genotype, we know the phenotype of the individual: *DD* and *Dd* are tall, and *dd* is dwarf. When both alleles are the same (*DD* or *dd*), the individual is **homozygous** for the trait, or a **homozygote**; when the alleles are different (*Dd*), we use the terms **heterozygous** and **heterozygote**. These symbols and terms are used in **Figure 3.2** to describe the monohybrid cross, as discussed on page 39.

Punnett Squares

The genotypes and phenotypes resulting from combining gametes during fertilization can be easily visualized by constructing a diagram called a **Punnett square**, named after the person who first devised this approach, Reginald C. Punnett. **Figure 3.3** illustrates this method of analysis for our $F_1 \times F_1$ monohybrid cross. Each of the possible gametes is assigned a column or a row; the vertical columns represent those of the female parent, and the horizontal rows represent those of the male parent. After assigning the gametes to the rows and columns, we predict the new generation by entering the male and female gametic information into each box and thus producing every possible resulting genotype. By filling out the Punnett square, we are listing all possible random fertilization events. The genotypes and phenotypes of all potential offspring are ascertained by reading the combinations in the boxes.

The Punnett square method is particularly useful when you are first learning about genetics and how to solve genetics problems. Note the ease with which the 3:1 phenotypic ratio and the 1:2:1 genotypic ratio may be derived for the F_2 generation in Figure 3.3.

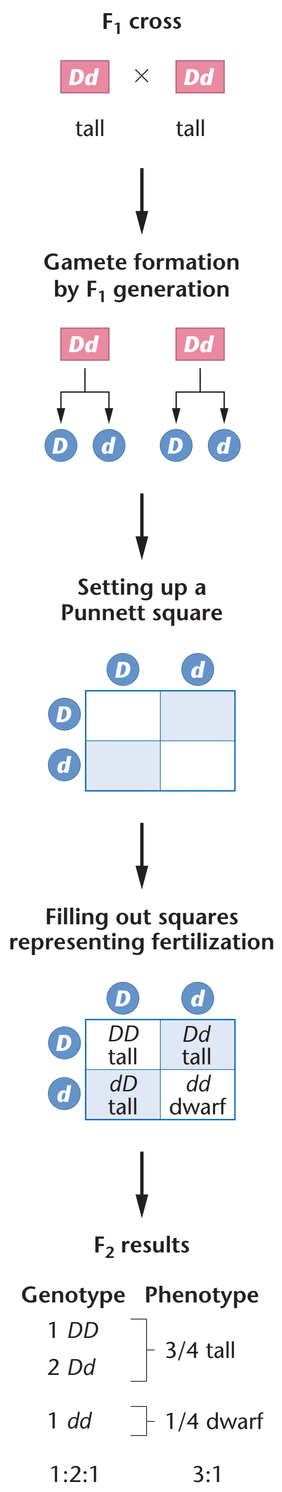


FIGURE 3.3 A Punnett square generating the F_2 ratio of the $F_1 \times F_1$ cross shown in Figure 3.2.

The Testcross: One Character

Tall plants produced in the F_2 generation are predicted to have either the *DD* or the *Dd* genotype. You might ask if there is a way to distinguish the genotype. Mendel devised a rather simple method that is still used today to discover the genotype of plants and animals: the **testcross**. The organism expressing the dominant phenotype but having an unknown genotype is crossed with a known *homozygous recessive individual*. For example, as shown in **Figure 3.4(a)**, if a tall plant of genotype *DD* is testcrossed with a dwarf plant, which must have the *dd* genotype, all offspring will be tall phenotypically and *Dd* genotypically. However, as shown in **Figure 3.4(b)**, if a tall plant is *Dd* and is crossed with a dwarf plant (*dd*), then one-half

NOW SOLVE THIS

3.1 Pigeons may exhibit a checkered or plain color pattern. In a series of controlled matings, the following data were obtained.

	F_1 Progeny	
P_1 Cross	Checkered	Plain
(a) checkered × checkered	36	0
(b) checkered × plain	38	0
(c) plain × plain	0	35

Then F_1 offspring were selectively mated with the following results. (The P_1 cross giving rise to each F_1 pigeon is indicated in parentheses.)

	F_2 Progeny	
$F_1 \times F_1$ Crosses	Checkered	Plain
(d) checkered (a) × plain (c)	34	0
(e) checkered (b) × plain (c)	17	14
(f) checkered (b) × checkered (b)	28	9
(g) checkered (a) × checkered (b)	39	0

How are the checkered and plain patterns inherited? Select and assign symbols for the genes involved, and determine the genotypes of the parents and offspring in each cross.

■ **HINT:** *This problem asks you to analyze the data produced from several crosses involving pigeons and to determine the mode of inheritance and the genotypes of the parents and offspring in a number of instances. The key to its solution is to first determine whether or not this is a monohybrid cross. To do so, convert the data to ratios that are characteristic of Mendelian crosses. In the case of this problem, ask first whether any of the F_2 ratios match Mendel's 3:1 monohybrid ratio. If so, the second step is to determine which trait is dominant and which is recessive.*

of the offspring will be tall (*Dd*) and the other half will be dwarf (*dd*). Therefore, a 1:1 tall/dwarf ratio demonstrates the heterozygous nature of the tall plant of unknown genotype. The results of the testcross reinforced Mendel's conclusion that separate unit factors control traits.

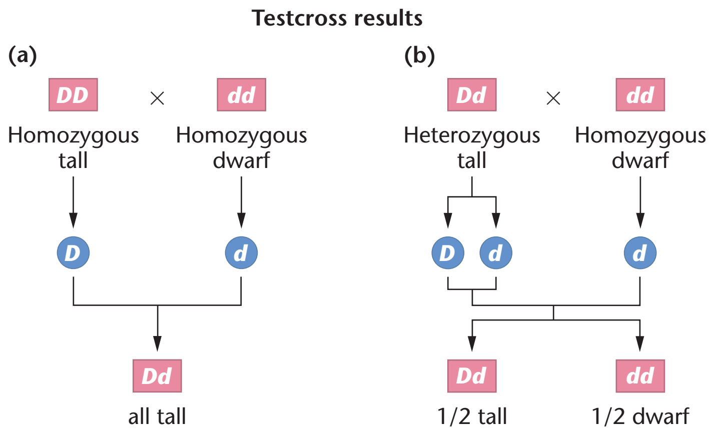


FIGURE 3.4 Testcross of a single character. In (a), the tall parent is homozygous, but in (b), the tall parent is heterozygous. The genotype of each tall P_1 plant can be determined by examining the offspring when each is crossed with the homozygous recessive dwarf plant.

3.3 Mendel's Dihybrid Cross Generated a Unique F_2 Ratio

As a natural extension of the monohybrid cross, Mendel also designed experiments in which he examined two characters simultaneously. Such a cross, involving two pairs of contrasting traits, is a **dihybrid cross**, or a *two-factor cross*. For example, if pea plants having yellow seeds that are round were bred with those having green seeds that are wrinkled, the results shown in **Figure 3.5** would occur: the F_1 offspring would all be yellow and round. It is therefore apparent that yellow is dominant to green and that round is dominant to wrinkled. When the F_1 individuals are selfed, approximately 9/16 of the F_2 plants express the yellow and round traits, 3/16 express yellow and wrinkled, 3/16 express green and round, and 1/16 express green and wrinkled.

A variation of this cross is also shown in Figure 3.5. Instead of crossing one P_1 parent with both dominant traits (yellow, round) to one with both recessive traits (green, wrinkled), plants with yellow, wrinkled seeds are crossed with those with green, round seeds. In spite of the change in the P_1 phenotypes, both the F_1 and F_2 results remain unchanged. Why this is so will become clear below.

Mendel's Fourth Postulate: Independent Assortment

We can most easily understand the results of a dihybrid cross if we consider it theoretically as consisting of two monohybrid crosses conducted separately. Think of the two sets of traits as being inherited independently of each other; that is, the chance of any plant having yellow or green seeds is not at all influenced by the chance that this plant will have round or wrinkled seeds. Thus, because yellow is dominant to green, all F_1 plants in the first theoretical cross would have yellow seeds. In the second theoretical cross, all F_1 plants would have round seeds because round is dominant to wrinkled. When Mendel examined the F_1 plants of the dihybrid cross, all were yellow and round, as our theoretical crosses predict.

The predicted F_2 results of the first cross are 3/4 yellow and 1/4 green. Similarly, the second cross would yield 3/4 round and 1/4 wrinkled. Figure 3.5 shows that in the dihybrid cross, 12/16 F_2 plants are yellow, while 4/16 are green, exhibiting the expected 3:1 (3/4:1/4) ratio. Similarly, 12/16 of all F_2 plants have round seeds, while 4/16 have wrinkled seeds, again revealing the 3:1 ratio.

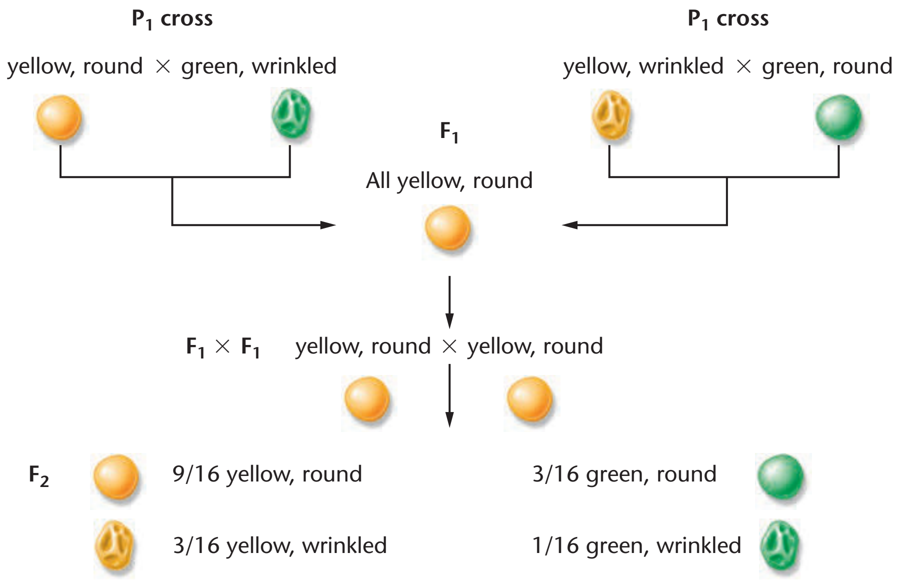


FIGURE 3.5 F_1 and F_2 results of Mendel's dihybrid crosses in which the plants on the top left with yellow, round seeds are crossed with plants having green, wrinkled seeds, and the plants on the top right with yellow, wrinkled seeds are crossed with plants having green, round seeds.

These numbers demonstrate that the two pairs of contrasting traits are inherited independently, so we can predict the frequencies of all possible F_2 phenotypes by applying the **product law** of probabilities: *the probability of two or more independent events occurring simultaneously is equal to the product of their individual probabilities.* For example, the probability of an F_2 plant having yellow and round seeds is (3/4)(3/4), or 9/16, because 3/4 of all F_2 plants should be yellow and 3/4 of all F_2 plants should be round.

In a like manner, the probabilities of the other three F_2 phenotypes can be calculated: yellow (3/4) and wrinkled (1/4) are predicted to be present together 3/16 of the time; green (1/4) and round (3/4) are predicted 3/16 of the time; and green (1/4) and wrinkled (1/4) are predicted 1/16 of the time. These calculations are shown in Figure 3.6.

It is now apparent why the F_1 and F_2 results are identical whether the initial cross is yellow, round plants bred with green, wrinkled plants, or whether yellow, wrinkled plants are bred with green, round plants. In both crosses, the F_1 genotype of all offspring is identical. As a result, the F_2 generation is also identical in both crosses.

On the basis of similar results in numerous dihybrid crosses, Mendel proposed a fourth postulate:

4. INDEPENDENT ASSORTMENT

   During gamete formation, segregating pairs of unit factors assort independently of each other.

   This postulate stipulates that segregation of any pair of unit factors occurs independently of all others. As a result of random segregation, each gamete receives one member of every pair of unit factors. For one pair, whichever unit factor is received does not influence the outcome of segregation of any other pair. Thus, according to the postulate of independent assortment, all possible combinations of gametes should be formed in equal frequency.

The Punnett square in Figure 3.7 shows how independent assortment works in the formation of the F_2 generation. Examine the formation of gametes by the F_1 plants; segregation prescribes that every gamete receives either a *G* or *g* allele and a *W* or *w* allele. Independent assortment stipulates that all four combinations (*GW*, *Gw*, *gW*, and *gw*) will be formed with equal probabilities.

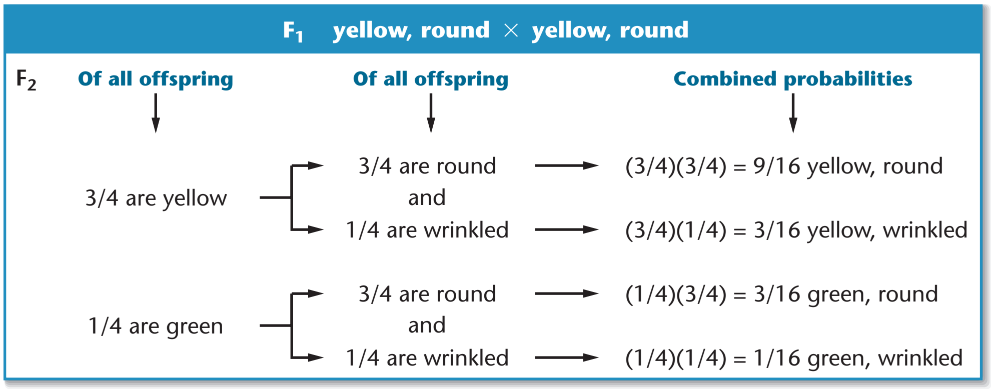
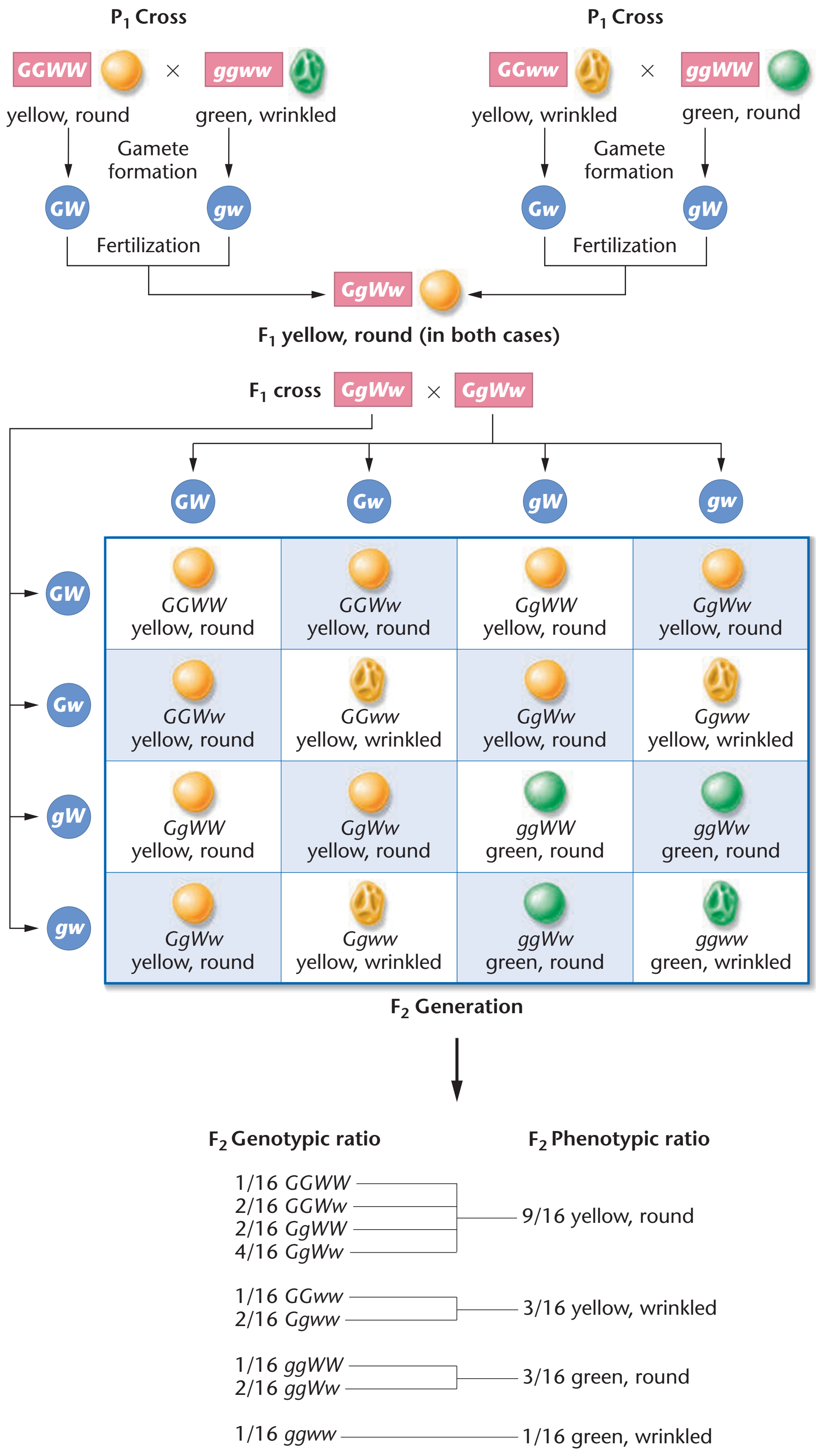

F_1 yellow, round × yellow, round
F_2
Of all offspring
Of all offspring
Combined probabilities
3/4 are yellow
3/4 are round
and
1/4 are wrinkled
(3/4)(3/4) = 9/16 yellow, round
(3/4)(1/4) = 3/16 yellow, wrinkled
1/4 are green
3/4 are round
and
1/4 are wrinkled
(1/4)(3/4) = 3/16 green, round
(1/4)(1/4) = 1/16 green, wrinkled

FIGURE 3.6 Computation of the combined probabilities of each F_2 phenotype for two independently inherited characters. The probability of each plant being yellow or green is independent of the probability of it bearing round or wrinkled seeds.



	GW	*Gw*	*gW*	*gw*
GW	*GGWW* yellow, round	*GGWw* yellow, round	*GgWW* yellow, round	*GgWw* yellow, round
Gw	*GGWw* yellow, round	*GGww* yellow, wrinkled	*GgWw* yellow, round	*Ggww* yellow, wrinkled
gW	*GgWW* yellow, round	*GgWw* yellow, round	*ggWW* green, round	*ggWw* green, round
gw	*GgWw* yellow, round	*Ggww* yellow, wrinkled	*ggWw* green, round	*ggww* green, wrinkled

F$_2$ Genotypic ratio	F$_2$ Phenotypic ratio
1/16 *GGWW* 2/16 *GGWw* 2/16 *GgWW* 4/16 *GgWw*	9/16 yellow, round
1/16 *GGww* 2/16 *Ggww*	3/16 yellow, wrinkled
1/16 *ggWW* 2/16 *ggWw*	3/16 green, round
1/16 *ggww*	1/16 green, wrinkled

FIGURE 3.7 Analysis of the dihybrid crosses shown in Figure 3.5. The F$_1$ heterozygous plants are self-fertilized to produce an F$_2$ generation, which is computed using a Punnett square. Both the phenotypic and genotypic F$_2$ ratios are shown.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Identifying Mendel's Gene for Regulating White Flower Color in Peas

In 2010, 150 years after Gregor Mendel studied pea flower color, an international team of researchers identified the gene responsible for regulating flower color in peas. The potential gene they focused on was called pea gene *A*. This gene was also found in other plants, including petunias and barrel clovers. Gene *A* encodes a protein that functions as a **transcription factor**—a protein that binds to DNA and regulates expression of other genes. Cells in purple flowers of pea plants accumulate anthocyanin pigment molecules that are responsible for their color. Pea plants with white flowers do not accumulate anthocyanin, even though they contain the gene that encodes the enzyme involved in anthocyanin synthesis. Researchers hypothesized that the transcription factor produced by pea gene *A* might regulate expression of the anthocyanin biosynthetic gene.

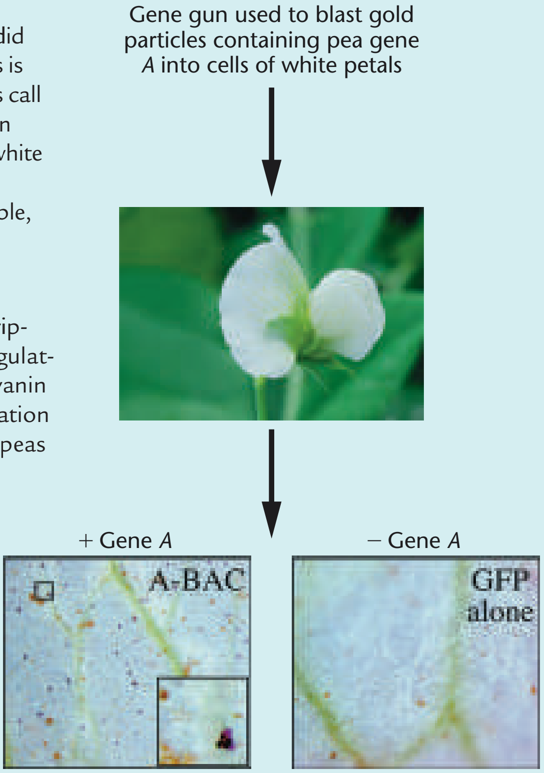
To test this hypothesis and confirm gene *A* function, they delivered normal copies of gene *A* into white flower petals by using a gene gun, a device that shoots metal particles coated with a gene of interest into cells. In this approach, gold particles coated with gene *A* enter a small percentage of cells and gene *A* is expressed in those cells.

Results:

Cells of white petals where gene *A* is expressed (left photo) accumulate anthocyanin pigment and turn purple. The inset square shows a higher magnification image enlarging spots of gene *A* expression. A control experiment where white petals were treated with DNA without gene *A* (right photo) did not restore pigmentation. This is an example of what geneticists call a **rescue experiment** because in cells that received gene *A* the white flower mutant phenotype was rescued or restored to the purple, wild-type phenotype.

Conclusion:

Pea gene *A* encodes a transcription factor responsible for regulating expression of the anthocyanin gene in peas. Further examination of the gene *A* sequence from peas with white flowers revealed a single-nucleotide change in gene *A* that renders this transcription factor inactive. Cells with a normal copy of gene *A* express anthocyanin and turn purple. Cells with a mutant form of gene *A* do not accumulate anthocyanin and are white. The genetic mystery of Mendel's white flowers had been solved.


Gene gun used to blast gold particles containing pea gene *A* into cells of white petals

Reference:

Hellens, R. P., et al. (2010). Identification of Mendel's White Flower Character. *PLoS One* 5(10): e13230. doi:10.1371/journal.pone.0013230. Take a look at the YouTube video "Finding the molecular answer to Mendel's pea colour experiments" to hear Dr. Hellens describe the approach his group used to identify the function of this gene. http://www.youtube.com/watch?v=BEhtyXCdcTg

Questions to Consider:

1. Why do you think that expression of gene *A* appears as spots in the leaves shown in the photo on the left? What does this signify?
2. If you did the same experiment with a pea plant that had a mutation in the gene for anthocyanin accumulation, would you expect that introduction of gene *A* would rescue the phenotype of the mutant? Why or why not?

In every $F_1 \times F_1$ fertilization event, each zygote has an equal probability of receiving one of the four combinations from each parent. If many offspring are produced, 9/16 have yellow, round seeds, 3/16 have yellow, wrinkled seeds, 3/16 have green, round seeds, and 1/16 have green, wrinkled seeds, yielding what is designated as **Mendel's 9:3:3:1 dihybrid ratio**. This is an ideal ratio based on probability events involving segregation, independent assortment, and random fertilization. Because of deviation due strictly to chance, particularly if small numbers of offspring are produced, actual results are highly unlikely to match the ideal ratio.

The Testcross: Two Characters

The testcross may also be applied to individuals that express two dominant traits but whose genotypes are unknown. For example, the expression of the yellow, round seed phenotype in the F_2 generation just described may result from the *GGWW, GGWw, GgWW*, or *GgWw* genotypes. If an F_2 yellow,

NOW SOLVE THIS

3.2 Considering the Mendelian traits round versus wrinkled and yellow versus green, consider the crosses below and determine the genotypes of the parental plants by analyzing the phenotypes of their offspring.

Parental Plants	Offspring
(a) round, yellow × round, yellow	3/4 round, yellow 1/4 wrinkled, yellow
(b) wrinkled, yellow × round, yellow	6/16 wrinkled, yellow 2/16 wrinkled, green 6/16 round, yellow 2/16 round, green
(c) round, yellow × round, yellow	9/16 round, yellow 3/16 round, green 3/16 wrinkled, yellow 1/16 wrinkled, green
(d) round, yellow × wrinkled, green	1/4 round, yellow 1/4 round, green 1/4 wrinkled, yellow 1/4 wrinkled, green

■ **HINT:** *This problem involves a series of Mendelian dihybrid crosses where you are asked to determine the genotypes of the parents in a number of instances. The key to its solution is to write down everything that you know for certain. This reduces the problem to its bare essentials, clarifying what you need to determine. For example, the wrinkled, yellow plant in case (b) must be homozygous for the recessive wrinkled alleles and bear at least one dominant allele for the yellow trait. Having established this, you need only determine the remaining allele for seed color.*

round plant is crossed with the homozygous recessive green, wrinkled plant (*ggww*), analysis of the offspring will indicate the exact genotype of that yellow, round plant. Each of the above genotypes results in a different set of gametes and, in a testcross, a different set of phenotypes in the resulting offspring. You should work out the results of each of these four crosses to be sure that you understand this concept.

3.4 The Trihybrid Cross Demonstrates That Mendel's Principles Apply to Inheritance of Multiple Traits

Thus far, we have considered inheritance of up to two pairs of contrasting traits. Mendel demonstrated that the processes of segregation and independent assortment also apply to three pairs of contrasting traits, in what is called a **trihybrid cross**, or *three-factor cross.*

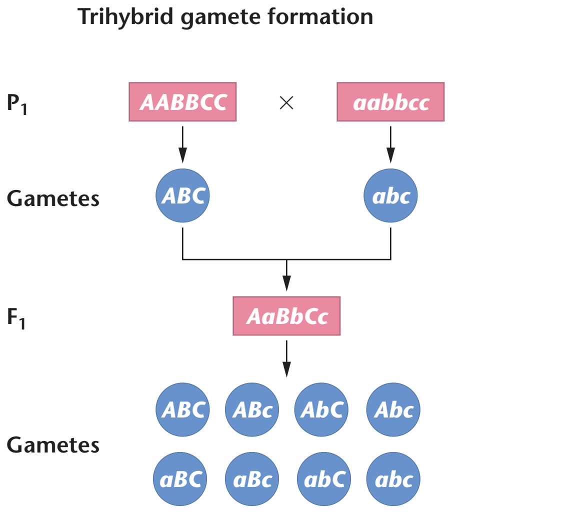


FIGURE 3.8 Formation of P_1 and F_1 gametes in a trihybrid cross.

Although a trihybrid cross is somewhat more complex than a dihybrid cross, its results are easily calculated if the principles of segregation and independent assortment are followed. For example, consider the cross shown in Figure 3.8 where the allele pairs of theoretical contrasting traits are represented by the symbols *A, a, B, b, C,* and *c*. In the cross between *AABBCC* and *aabbcc* individuals, all F_1 individuals are heterozygous for all three gene pairs. Their genotype, *AaBbCc*, results in the phenotypic expression of the dominant *A, B,* and *C* traits. When F_1 individuals serve as parents, each produces eight different gametes in equal frequencies. At this point, we could construct a Punnett square with 64 separate boxes and read out the phenotypes—but such a method is cumbersome in a cross involving so many factors. Therefore, another method has been devised to calculate the predicted ratio.

The Forked-Line Method, or Branch Diagram

It is much less difficult to consider each contrasting pair of traits separately and then to combine these results by using the **forked-line method**, first shown in Figure 3.6. This method, also called a **branch diagram**, relies on the simple application of the laws of probability established for the dihybrid cross. Each gene pair is assumed to behave independently during gamete formation.

When the monohybrid cross *AA* × *aa* is made, we know that:

1. All F_1 individuals have the genotype *Aa* and express the phenotype represented by the *A* allele, which is called the *A* phenotype in the discussion that follows.
2. The F_2 generation consists of individuals with either the *A* phenotype or the *a* phenotype in the ratio of 3:1.

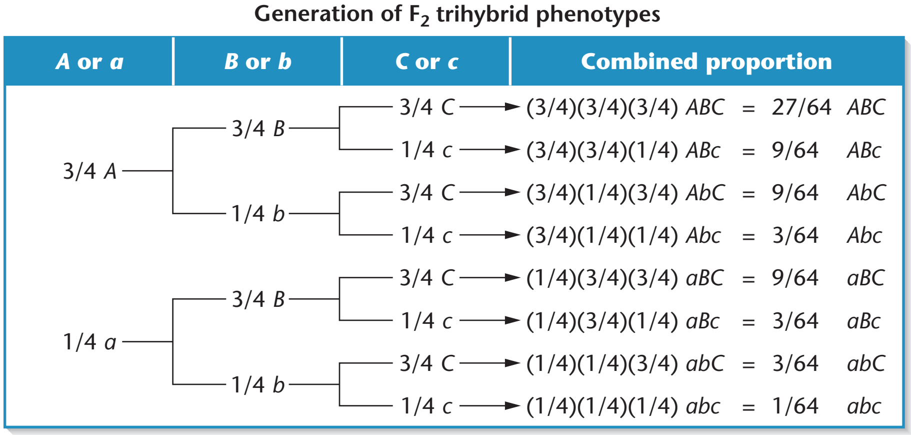

Generation of F_2 trihybrid phenotypes
A or a
B or b
C or c
Combined proportion
3/4 A
3/4 B
3/4 C → (3/4)(3/4)(3/4) ABC = 27/64 ABC
1/4 c → (3/4)(3/4)(1/4) ABc = 9/64 ABc
1/4 b
3/4 C → (3/4)(1/4)(3/4) AbC = 9/64 AbC
1/4 c → (3/4)(1/4)(1/4) Abc = 3/64 Abc
1/4 a
3/4 B
3/4 C → (1/4)(3/4)(3/4) aBC = 9/64 aBC
1/4 c → (1/4)(3/4)(1/4) aBc = 3/64 aBc
1/4 b
3/4 C → (1/4)(1/4)(3/4) abC = 3/64 abC
1/4 c → (1/4)(1/4)(1/4) abc = 1/64 abc

FIGURE 3.9 Generation of the F_2 trihybrid phenotypic ratio using the forked-line method. This method is based on the expected probability of occurrence of each phenotype.

The same generalizations can be made for the *BB* × *bb* and *CC* × *cc* crosses. Thus, in the F_2 generation, 3/4 of all organisms will express phenotype *A*, 3/4 will express *B*, and 3/4 will express *C*. Similarly, 1/4 of all organisms will express *a*, 1/4 will express *b*, and 1/4 will express *c*. The proportions of organisms that express each phenotypic combination can be predicted by assuming that fertilization, following the independent assortment of these three gene pairs during gamete formation, is a random process. We apply the product law of probabilities once again. **Figure 3.9** uses the forked-line method to calculate the phenotypic proportions of the F_2 generation. They fall into the trihybrid ratio of 27:9:9:9:3:3:3:1. The same method can be used to solve crosses involving any number of gene pairs, *provided that all gene pairs assort independently from each other.* We shall see later that gene pairs do not always assort with complete independence. However, it appeared to be true for all of Mendel's characters.

NOW SOLVE THIS

3.3 Using the forked-line, or branch diagram, method, determine the genotypic and phenotypic ratios of these trihybrid crosses: (a) *AaBbCc* × *AaBBCC*, (b) *AaBBCc* × *aaBBCc*, and (c) *AaBbCc* × *AaBbCc*.

■ **HINT:** *This problem asks you to use the forked-line method to determine the outcome of a number of trihybrid crosses. The key to its solution is to realize that in using the forked-line method, you must consider each gene pair separately. For example, in this problem, first predict the outcome of each cross for the A/a genes, then for the B/b genes, and finally, for the C/c genes. Then you are prepared to pursue the outcome of each cross using the forked-line method.*

3.5 Mendel's Work Was Rediscovered in the Early Twentieth Century

Mendel published his work in 1866. While his findings were often cited and discussed, their significance went unappreciated for about 35 years. Then, in the latter part of the nineteenth century, a remarkable observation set the scene for the recognition of Mendel's work: Walter Flemming's discovery of chromosomes in the nuclei of salamander cells. In 1879, Flemming described the behavior of these thread-like structures during cell division. As a result of his findings and the work of many other cytologists, the presence of discrete units within the nucleus soon became an integral part of scientists' ideas about inheritance.

In the early twentieth century, hybridization experiments similar to Mendel's were performed independently by three botanists, Hugo de Vries, Carl Correns, and Erich Tschermak. De Vries's work demonstrated the principle of segregation in several plant species. Apparently, he searched the existing literature and found that Mendel's work had anticipated his own conclusions! Correns and Tschermak also reached conclusions similar to those of Mendel.

About the same time, two cytologists, Walter Sutton and Theodor Boveri, independently published papers linking their discoveries of the behavior of chromosomes during meiosis to the Mendelian principles of segregation and independent assortment. They pointed out that the separation of chromosomes during meiosis could serve as the cytological basis of these two postulates. Although they thought that Mendel's unit factors were probably chromosomes rather than genes on chromosomes, their findings reestablished the importance of Mendel's work and led to many ensuing genetic investigations. Sutton and Boveri are credited with initiating the **chromosomal theory of inheritance**, the idea that the genetic material in living organisms is contained in chromosomes, which was developed during the next two decades. As we will see in subsequent chapters, work by Thomas H. Morgan, Alfred H. Sturtevant, Calvin Bridges, and others established beyond a reasonable doubt that Sutton's and Boveri's hypothesis was correct.

Unit Factors, Genes, and Homologous Chromosomes

Because the correlation between Sutton's and Boveri's observations and Mendelian postulates serves as the foundation for the modern description of transmission genetics, we will examine this correlation in some depth before moving on to other topics.

As we know, each species possesses a specific number of chromosomes in each somatic cell nucleus. For diploid organisms, this number is called the **diploid number (2*n*)** and is characteristic of that species. During the formation of gametes (meiosis), the number is precisely halved (*n*), and when two gametes combine during fertilization, the diploid number is reestablished. During meiosis, however, the chromosome number is not reduced in a random manner. It was apparent to early cytologists that the diploid number of chromosomes is composed of homologous pairs identifiable by their morphological appearance and behavior. The gametes contain one member of each pair—thus the chromosome complement of a gamete is quite specific, and the number of chromosomes in each gamete is equal to the haploid number.

With this basic information, we can see the correlation between the behavior of unit factors and chromosomes and genes. **Figure 3.10** shows three of Mendel's postulates and

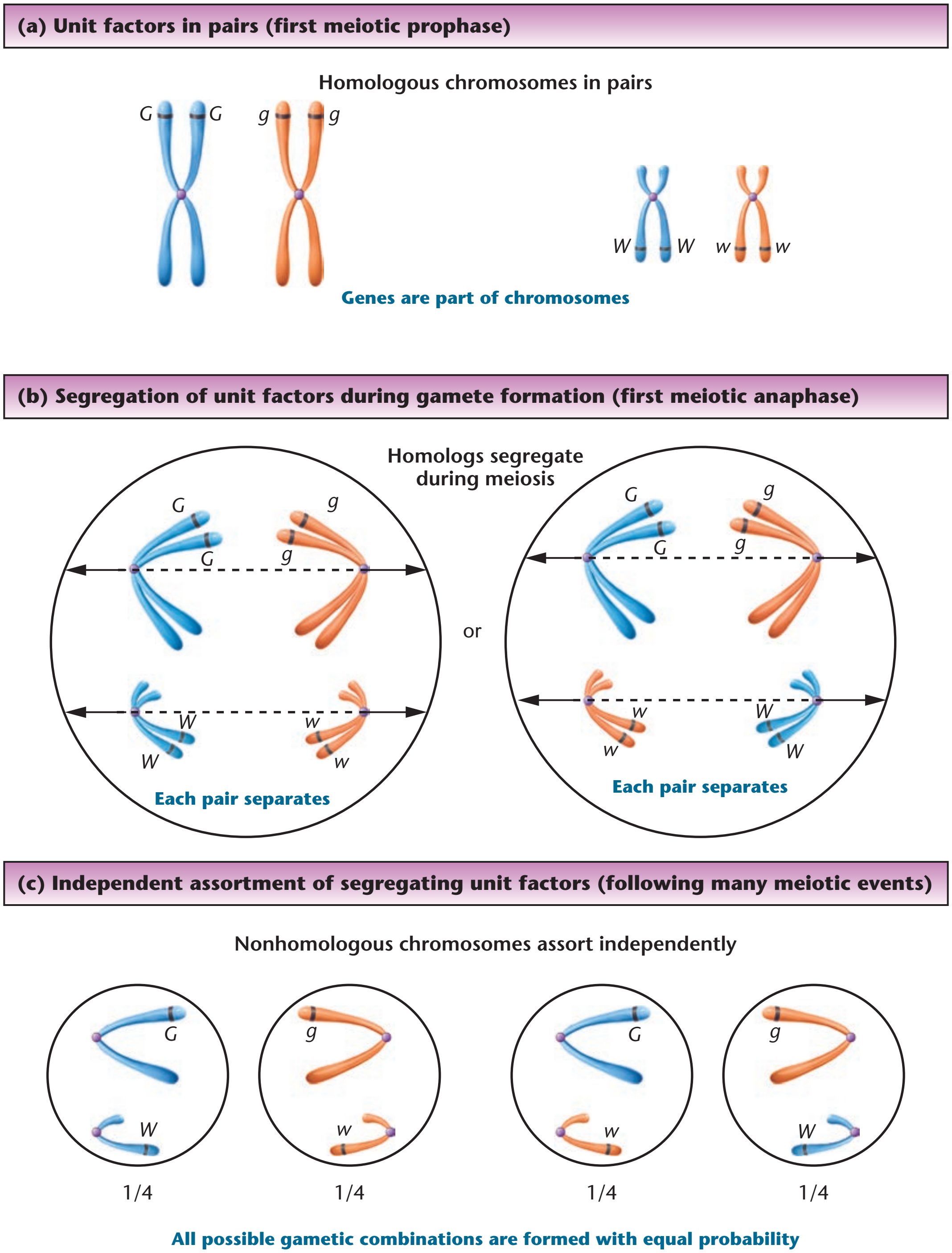

(a) Unit factors in pairs (first meiotic prophase)
Homologous chromosomes in pairs
G
G
g
g
W
W
w
w
Genes are part of chromosomes
(b) Segregation of unit factors during gamete formation (first meiotic anaphase)
Homologs segregate during meiosis
G
g
G
g
W
w
W
w
or
G
g
G
g
w
W
w
W
Each pair separates
Each pair separates
(c) Independent assortment of segregating unit factors (following many meiotic events)
Nonhomologous chromosomes assort independently
G
W
g
w
G
w
g
W
1/4
1/4
1/4
1/4
All possible gametic combinations are formed with equal probability

FIGURE 3.10 Illustrated correlation between the Mendelian postulates of (a) unit factors in pairs, (b) segregation, and (c) independent assortment, showing the presence of genes located on homologous chromosomes and their behavior during meiosis.

the chromosomal explanation of each. Unit factors are really genes located on homologous pairs of chromosomes [Figure 3.10(a)]. Members of each pair of homologs separate, or segregate, during gamete formation [Figure 3.10(b)]. In the figure, two different alignments are possible, both of which are shown.

To illustrate the principle of independent assortment, it is important to distinguish between members of any given homologous pair of chromosomes. One member of each pair is derived from the **maternal parent**, whereas the other comes from the **paternal parent**. (We represent the different parental origins with different colors.) As shown in Figure 3.10(c), following independent segregation of each pair of homologs, each gamete receives one member from each pair of chromosomes. All possible combinations are formed with equal probability. If we add the symbols used in Mendel's dihybrid cross (*G, g* and *W, w*) to the diagram, we can see why equal numbers of the four types of gametes are formed. The independent behavior of Mendel's pairs of unit factors (*G* and *W* in this example) is due to their presence on separate pairs of homologous chromosomes.

Observations of the phenotypic diversity of living organisms make it logical to assume that there are many more genes than chromosomes. Therefore, each homolog must carry genetic information for more than one trait. The currently accepted concept is that a chromosome is composed of a large number of linearly ordered, information-containing genes. Mendel's paired unit factors (which determine tall or dwarf stems, for example) actually constitute a pair of genes located on one pair of homologous chromosomes. The location on a given chromosome where any particular gene occurs is called its **locus** (pl. loci). The different alleles of a given gene (for example, *G* and *g*) contain slightly different genetic information (green or yellow) that determines the same character (seed color in this case). Although we have examined only genes with two alternative alleles, most genes have more than two allelic forms. We conclude this section by reviewing the criteria necessary to classify two chromosomes as a homologous pair:

1. During mitosis and meiosis, when chromosomes are visible in their characteristic shapes, both members of a homologous pair are the same size and exhibit identical centromere locations. The sex chromosomes (e.g., the X and the Y chromosomes in mammals) are an exception.
2. During early stages of meiosis, homologous chromosomes form pairs, or synapse.
3. Although it is not generally visible under the microscope, homologs contain the identical linear order of gene loci.

EVOLVING CONCEPT OF THE GENE

Based on the pioneering work of Gregor Mendel, the gene was viewed as a heritable unit factor that determines the expression of an observable trait, or phenotype. ■

3.6 Independent Assortment Leads to Extensive Genetic Variation

One consequence of independent assortment is the production by an individual of genetically dissimilar gametes. Genetic variation results because the two members of any homologous pair of chromosomes are rarely, if ever, genetically identical. As the maternal and paternal members of all pairs are distributed to gametes through independent assortment, all possible chromosome combinations are produced, leading to extensive genetic diversity.

We have seen that the number of possible gametes, each with different chromosome compositions, is 2^n, where n equals the haploid number. Thus, if a species has a haploid number of 4, then 2^4, or 16, different gamete combinations can be formed as a result of independent assortment. Although this number is not high, consider the human species, where $n = 23$. When 2^{23} is calculated, we find that in excess of 8×10^6, or over 8 million, different types of gametes are possible through independent assortment. Because fertilization represents an event involving only one of approximately 8×10^6 possible gametes from each of two parents, each offspring represents only one of $(8 \times 10^6)^2$ or one of only 64×10^{12} potential genetic combinations. Given that this probability is less than one in one trillion, it is no wonder that, except for identical twins, each member of the human species exhibits a distinctive set of traits—this number of combinations of chromosomes is far greater than the number of humans who have ever lived on Earth! Genetic variation resulting from independent assortment has been extremely important to the process of evolution in all sexually reproducing organisms.

3.7 Laws of Probability Help to Explain Genetic Events

Recall that genetic ratios—for example, 3/4 tall:1/4 dwarf—are most properly thought of as probabilities. These values predict the outcome of each fertilization event, such that the probability of each zygote having the genetic potential for becoming tall is 3/4, whereas the potential for its being a dwarf is 1/4. Probabilities range from 0.0, where an event *is certain not to occur,* to 1.0, where an event *is certain to occur.* In this section, we consider the relation of probability to genetics. When two or more events with known probabilities occur independently but at the same time, we can calculate the probability of their possible outcomes occurring together. This is accomplished by applying the **product law**, which states that *the probability of two or more independent events occurring simultaneously is equal to the product of their individual probabilities* (see Section 3.3). Two or more events are independent of one another if the outcome of each one does not affect the outcome of any of the others under consideration.

To illustrate the product law, consider the possible results if you toss a penny (P) and a nickel (N) at the same time and examine all combinations of heads (H) and tails (T) that can occur. There are four possible outcomes:

$$\begin{aligned}(P_H{:}N_H) &= (1/2)(1/2) = 1/4\\(P_T{:}N_H) &= (1/2)(1/2) = 1/4\\(P_H{:}N_T) &= (1/2)(1/2) = 1/4\\(P_T{:}N_T) &= (1/2)(1/2) = 1/4\end{aligned}$$

The probability of obtaining a head or a tail in the toss of either coin is 1/2 and is unrelated to the outcome for the other coin. Thus, all four possible combinations are predicted to occur with equal probability.

If we want to calculate the probability when the possible outcomes of two events are independent of one another but can be accomplished in more than one way, we can apply the **sum law**. For example, what is the probability of tossing our penny and nickel and obtaining one head and one tail? In such a case, we do not care whether it is the penny or the nickel that comes up heads, provided that the other coin has the alternative outcome. As we saw above, there are two ways in which the desired outcome can be accomplished, each with a probability of 1/4. The sum law states that *the probability of obtaining any single outcome, where that outcome can be achieved by two or more events, is equal to the sum of the individual probabilities of all such events.* Thus, according to the sum law, the overall probability in our example is equal to

$$(1/4) + (1/4) = 1/2$$

One-half of all two-coin tosses are predicted to yield the desired outcome.

These simple probability laws will be useful throughout our discussions of transmission genetics and for solving genetics problems. In fact, we already applied the product law when we used the forked-line method to calculate the phenotypic results of Mendel's dihybrid and trihybrid crosses. When we wish to know the results of a cross, we need only calculate the probability of each possible outcome. The results of this calculation then allow us to predict the proportion of offspring expressing each phenotype or each genotype.

An important point to remember when you deal with probability is that predictions of possible outcomes are based on large sample sizes. If we predict that 9/16 of the offspring of a dihybrid cross will express both dominant traits, it is very unlikely that, in a small sample, exactly 9 of every 16 will express this phenotype. Instead, our prediction is that, of a large number of offspring, approximately 9/16 will do so. The deviation from the predicted ratio in smaller sample sizes is attributed to chance, a subject we examine in our discussion of statistics in Section 3.8. As you shall see, the impact of deviation due strictly to chance diminishes as the sample size increases.

3.8 Chi-Square Analysis Evaluates the Influence of Chance on Genetic Data

Mendel's 3:1 monohybrid and 9:3:3:1 dihybrid ratios are hypothetical predictions based on the following assumptions: (1) each allele is dominant or recessive, (2) segregation is unimpeded, (3) independent assortment occurs, and (4) fertilization is random. The final two assumptions are influenced by chance events and therefore are subject to random fluctuation. This concept of **chance deviation** is most easily illustrated by tossing a single coin numerous times and recording the number of heads and tails observed. In each toss, there is a probability of 1/2 that a head will occur and a probability of 1/2 that a tail will occur. Therefore, the expected ratio of many tosses is 1/2:1/2, or 1:1. If a coin is tossed 1000 times, usually *about* 500 heads and 500 tails will be observed. Any reasonable fluctuation from this hypothetical ratio (e.g., 486 heads and 514 tails) is attributed to chance.

As the total number of tosses is reduced, the impact of chance deviation increases. For example, if a coin is tossed only four times, you would not be too surprised if all four tosses resulted in only heads or only tails. For 1000 tosses, however, 1000 heads or 1000 tails would be most unexpected. In fact, you might believe that such a result would be impossible. Actually, all heads or all tails in 1000 tosses can be predicted to occur with a probability of $(1/2)^{1000}$. Since $(1/2)^{20}$ is less than one in a million times, an event occurring with a probability as small as $(1/2)^{1000}$ is virtually impossible. Two major points to keep in mind when predicting or analyzing genetic outcomes are:

1. The outcomes of independent assortment and fertilization, like coin tossing, are subject to random fluctuations from their predicted occurrences as a result of chance deviation.
2. As the sample size increases, the average deviation from the expected results decreases. Therefore, a larger sample size diminishes the impact of chance deviation on the final outcome.

Chi-Square Calculations and the Null Hypothesis

In genetics, being able to evaluate observed deviation is a crucial skill. When we assume that data will fit a given ratio such as 1:1, 3:1, or 9:3:3:1, we establish what is called the **null hypothesis (H_0)**. It is so named because the hypothesis assumes that there is *no real difference* between the *measured values* (or ratio) and the *predicted values* (or ratio). Any apparent difference can be attributed purely to chance. The validity of the null hypothesis for a given set of data is measured using statistical analysis. Depending on the results of this analysis, the null hypothesis may either (1) *be rejected* or (2) *fail to be rejected*. If it is rejected, the observed deviation from the expected result is judged not to be attributable to chance alone. In this case, the null hypothesis and the underlying assumptions leading to it must be reexamined. If the null hypothesis fails to be rejected, any observed deviations are attributed to chance.

One of the simplest statistical tests for assessing the goodness of fit of the null hypothesis is **chi-square (χ^2) analysis**. This test takes into account the observed deviation in each component of a ratio (from what was expected) as well as the sample size and reduces them to a single numerical value. The value for χ^2 is then used to estimate how frequently the observed deviation can be expected to occur strictly as a result of chance. The formula used in chi-square analysis is

$$\chi^2 = \Sigma \frac{(o - e)^2}{e}$$

where o is the observed value for a given category, e is the expected value for that category, and Σ (the Greek letter sigma) represents the sum of the calculated values for each category in the ratio. Because $(o - e)$ is the deviation (d) in each case, the equation reduces to

$$\chi^2 = \Sigma \frac{d^2}{e}$$

Table 3.1(a) shows the steps in the χ^2 calculation for the F_2 results of a hypothetical monohybrid cross. To analyze the data obtained from this cross, work from left to right across the table, verifying the calculations as appropriate. Note that regardless of whether the deviation d is positive or negative, d^2 always becomes positive after the number is squared. In **Table 3.1(b)** F_2 results of a hypothetical dihybrid cross are analyzed. Make sure that you understand how each number was calculated in this example.

The final step in chi-square analysis is to interpret the χ^2 value. To do so, you must initially determine a value called the **degrees of freedom (*df*)**, which is equal to $n - 1$, where n is the number of different categories into which the data are divided, in other words, the number of possible outcomes. For the 3:1 ratio, $n = 2$, so $df = 1$. For the 9:3:3:1 ratio, $n = 4$ and $df = 3$. Degrees of freedom must be taken into account because the greater the number of categories, the more deviation is expected as a result of chance.

Once you have determined the degrees of freedom, you can interpret the χ^2 value in terms of a corresponding **probability value (*p*)**. Since this calculation is complex, we usually take the p value from a standard table or graph. **Figure 3.11** shows a wide range of χ^2 values and the corresponding p values for various degrees of freedom in both a graph and a table. Let's use the graph to explain how to determine the p value. The caption for Figure 3.11(b) explains how to use the table.

To determine p using the graph, execute the following steps:

1. Locate the χ^2 value on the abscissa (the horizontal axis, or x-axis).
2. Draw a vertical line from this point up to the line on the graph representing the appropriate *df*.

TABLE 3.1 Chi-Square Analysis

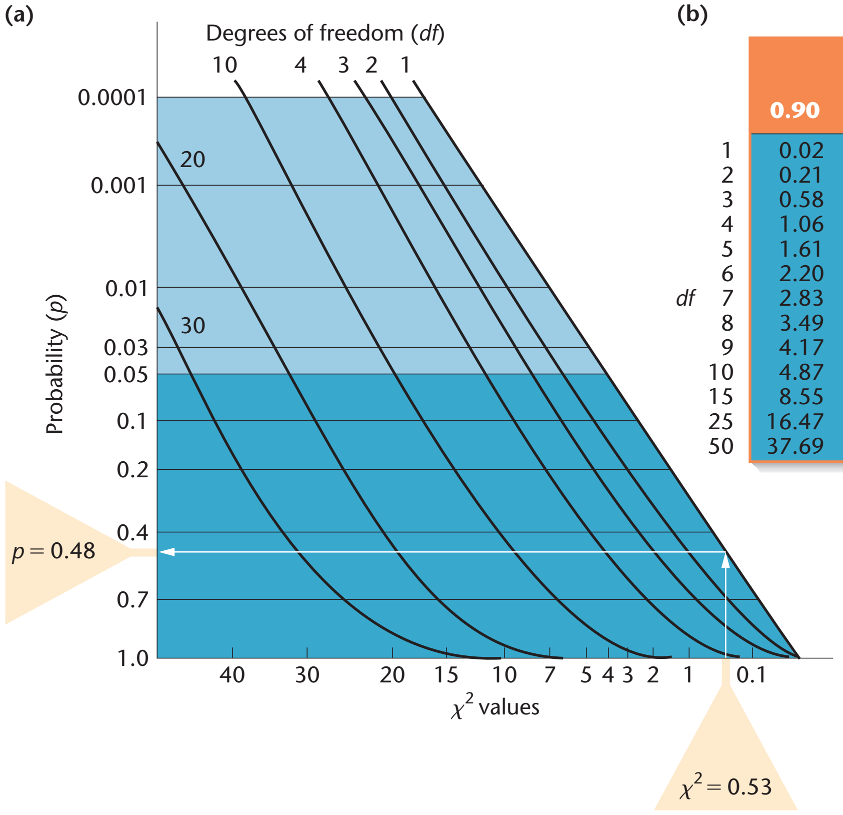
(a) Monohybrid

Cross Expected Ratio	Observed (o)	Expected (e)	Deviation ($o - e = d$)	Deviation^2	d^2/e
3/4	740	3/4(1000) = 750	740 − 750 = −10	$(-10)^2 = 100$	100/750 = 0.13
1/4	260	1/4(1000) = 250	260 − 250 = +10	$(+10)^2 = 100$	100/250 = 0.40
	Total = 1000				$\chi^2 = 0.53$
					$p = 0.48$

(b) Dihybrid

Cross Expected Ratio	Observed (o)	Expected (e)	Deviation ($o - e = d$)	Deviation^2	d^2/e
9/16	587	567	+20	400	0.71
3/16	197	189	+8	64	0.34
3/16	168	189	−21	441	2.33
1/16	56	63	−7	49	0.78
	Total = 1008				$\chi^2 = 4.16$
					$p = 0.26$

(a)



(b)

	Probability (*p*)					
df	**0.90**	0.50	0.20	**0.05**	**0.01**	**0.001**
1	0.02	0.46	1.64	3.84	6.64	10.83
2	0.21	1.39	3.22	5.99	9.21	13.82
3	0.58	2.37	4.64	7.82	11.35	16.27
4	1.06	3.36	5.99	9.49	13.28	18.47
5	1.61	4.35	7.29	11.07	15.09	20.52
6	2.20	5.35	8.56	12.59	16.81	22.46
7	2.83	6.35	9.80	14.07	18.48	24.32
8	3.49	7.34	11.03	15.51	20.09	26.13
9	4.17	8.34	12.24	16.92	21.67	27.88
10	4.87	9.34	13.44	18.31	23.21	29.59
15	8.55	14.34	19.31	25.00	30.58	37.30
25	16.47	24.34	30.68	37.65	44.31	52.62
50	37.69	49.34	58.16	67.51	76.15	86.60

χ^2 values

▇ Fails to reject the null hypothesis
▨ Rejects the null hypothesis

FIGURE 3.11 (a) Graph for converting χ^2 values to *p* values. (b) Table of χ^2 values for selected values of *df* and *p*. χ^2 values that lead to a *p* value of 0.05 or greater (darker blue areas) justify failure to reject the null hypothesis. Values leading to a *p* value of less than 0.05 (lighter blue areas) justify rejecting the null hypothesis. For example, the table in part (b) shows that for $\chi^2 = 0.53$ with 1 degree of freedom, the corresponding *p* value is between 0.20 and 0.50. The graph in (a) gives a more precise *p* value of 0.48 by interpolation. Thus, we fail to reject the null hypothesis.

3. From there, extend a horizontal line to the left until it intersects the ordinate (the vertical axis, or *y*-axis).
4. Estimate, by interpolation, the corresponding *p* value.

We used these steps for the monohybrid cross in Table 3.1(a) to estimate the *p* value of 0.48, as shown in Figure 3.11(a). Now try this method to see if you can determine the *p* value for the dihybrid cross [Table 3.1(b)]. Since the χ^2 value is 4.16 and $df = 3$, an approximate *p* value is 0.26. Checking this result in the table confirms that *p* values for both the monohybrid and dihybrid crosses are between 0.20 and 0.50.

Interpreting Probability Values

So far, we have been concerned with calculating χ^2 values and determining the corresponding *p* values. These steps bring us to the most important aspect of chi-square analysis: understanding the meaning of the *p* value. It is simplest to think of the *p* value as a percentage. Let's use the example of the dihybrid cross in Table 3.1(b) where $p = 0.26$, which can be thought of as 26 percent. In our example, the *p* value indicates that if we repeat the same experiment many times, 26 percent of the trials would be expected to exhibit chance deviation as great as or greater than that seen in the initial trial. Conversely, 74 percent of the repeats would show less deviation than initially observed as a result of chance. Thus, the *p* value reveals that a null hypothesis (concerning the 9:3:3:1 ratio, in this case) is never proved or disproved absolutely. Instead, a relative standard is set that we use to either *reject* or *fail to reject* the null hypothesis. This standard is most often a *p* value of 0.05. When applied to chi-square analysis, a *p* value less than 0.05 means that the observed deviation in the set of results will be obtained by chance alone less than 5 percent of the time. Such a *p* value indicates that the difference between the observed and predicted results is substantial and requires us to reject the null hypothesis.

On the other hand, *p* values of 0.05 or greater (0.05 to 1.0) indicate that the observed deviation will be obtained by chance alone 5 percent or more of the time. This conclusion allows us not to reject the null hypothesis (when we are using $p = 0.05$ as our standard). Thus, with its *p* value of 0.26, the null hypothesis that independent assortment accounts for the results fails to be rejected. Therefore, the observed deviation can be reasonably attributed to chance.

A final note is relevant here concerning the case where the null hypothesis is rejected, that is, where $p \leq 0.05$. Suppose we had tested a dataset to assess a possible 9:3:3:1 ratio, as in Table 3.1(b), but we rejected the null hypothesis based on our calculation. What are alternative interpretations of the data? Researchers will reassess the assumptions that underlie the null hypothesis. In our dyhibrid cross, we assumed that segregation operates faithfully for both gene pairs. We also assumed that fertilization is random and that the viability of all gametes is equal regardless of genotype—that is, all gametes are equally likely to participate in fertilization. Finally, we assumed that, following fertilization, all preadult stages and adult offspring are equally viable, regardless of their genotype. If any of these assumptions is incorrect, then the original hypothesis is not necessarily invalid.

An example will clarify this point. Suppose our null hypothesis is that a dihybrid cross between fruit flies will result in 3/16 mutant wingless flies. However, perhaps fewer of the mutant embryos are able to survive their preadult development or young adulthood compared to flies whose genotype gives rise to wings. As a result, when the data are gathered, there will be fewer than 3/16 wingless flies. Rejection of the null hypothesis is not in itself cause for us to reject the validity of the postulates of segregation and independent assortment, because other factors we are unaware of may also be affecting the outcome.

NOW SOLVE THIS

3.4 In one of Mendel's dihybrid crosses, he observed 315 round, yellow; 108 round, green; 101 wrinkled, yellow; and 32 wrinkled, green F_2 plants. Analyze these data using the χ^2 test to see if

(a) they fit a 9:3:3:1 ratio.
(b) the round:wrinkled data fit a 3:1 ratio.
(c) the yellow:green data fit a 3:1 ratio.

■ **HINT:** *This problem asks you to apply χ^2 analysis to a set of data and to determine whether those data fit any of several ratios. The key to its solution is to first calculate χ^2 by initially determining the expected outcomes using the predicted ratios. Then follow a stepwise approach, determining the deviation in each case, and calculating d^2/e for each category. Once you have determined the χ^2 value, you must then determine and interpret the p value for each ratio.*

For more practice, see Problems 18, 19, and 20.

3.9 Pedigrees Reveal Patterns of Inheritance of Human Traits

We now explore how to determine the mode of inheritance of phenotypes in humans, where experimental matings are not made and where relatively few offspring are available for study. The traditional way to study inheritance has been to construct a family tree, indicating the presence or absence of the trait in question for each member of each generation. Such a family tree is called a **pedigree**. By analyzing a pedigree, we may be able to predict how the trait under study is inherited—for example, is it due to a dominant or recessive allele? When many pedigrees for the same trait are studied, we can often ascertain the mode of inheritance.

Pedigree Conventions

Figure 3.12 illustrates some of the conventions geneticists follow in constructing pedigrees. Circles represent females and squares designate males. If the sex of an individual is unknown, a diamond is used. Parents are generally connected to each other by a single horizontal line, and vertical lines lead to their offspring. If the parents are related—that is, **consanguineous**—such as first cousins, they are connected by a double line. Offspring are called **sibs** (short for **siblings**) and are connected by a horizontal **sibship line**. Sibs are placed in birth order from left to right and are labeled with Arabic numerals. Parents also receive an Arabic number designation. Each generation is indicated by a Roman numeral. When a pedigree traces only a single trait, the circles, squares, and diamonds are shaded if the phenotype being considered is expressed and unshaded if not. In some pedigrees, those individuals that fail

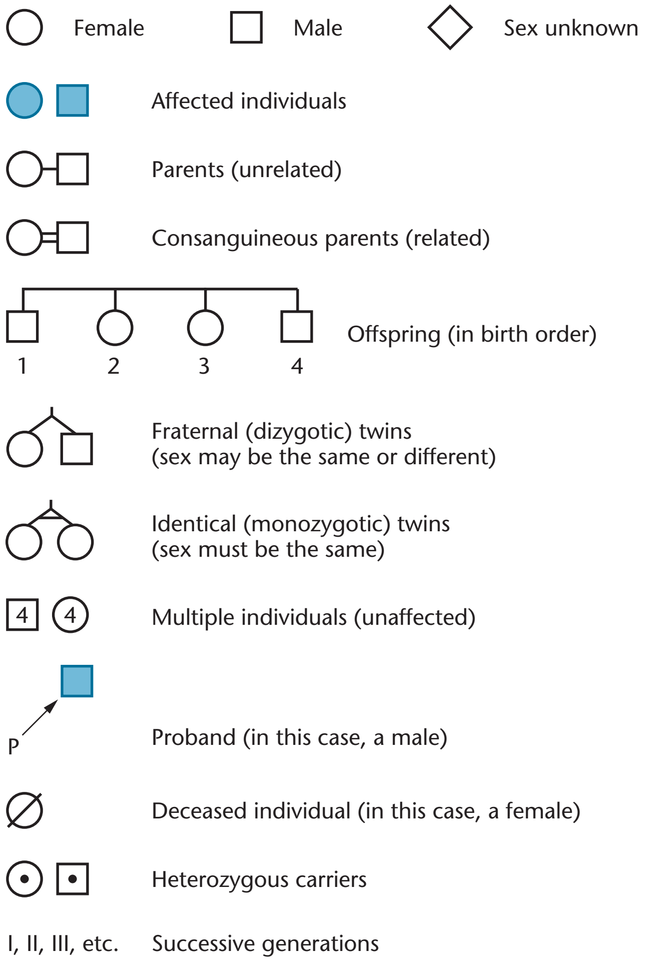

Female
Male
Sex unknown
Affected individuals
Parents (unrelated)
Consanguineous parents (related)
1
2
3
4
Offspring (in birth order)
Fraternal (dizygotic) twins
(sex may be the same or different)
Identical (monozygotic) twins
(sex must be the same)
4
4
Multiple individuals (unaffected)
P
Proband (in this case, a male)
Deceased individual (in this case, a female)
Heterozygous carriers
I, II, III, etc.
Successive generations

FIGURE 3.12 Conventions commonly encountered in human pedigrees.

to express a recessive trait but are known with certainty to be heterozygous carriers have a shaded dot within their unshaded circle or square. If an individual is deceased and the phenotype is unknown, a diagonal line is placed over the circle or square.

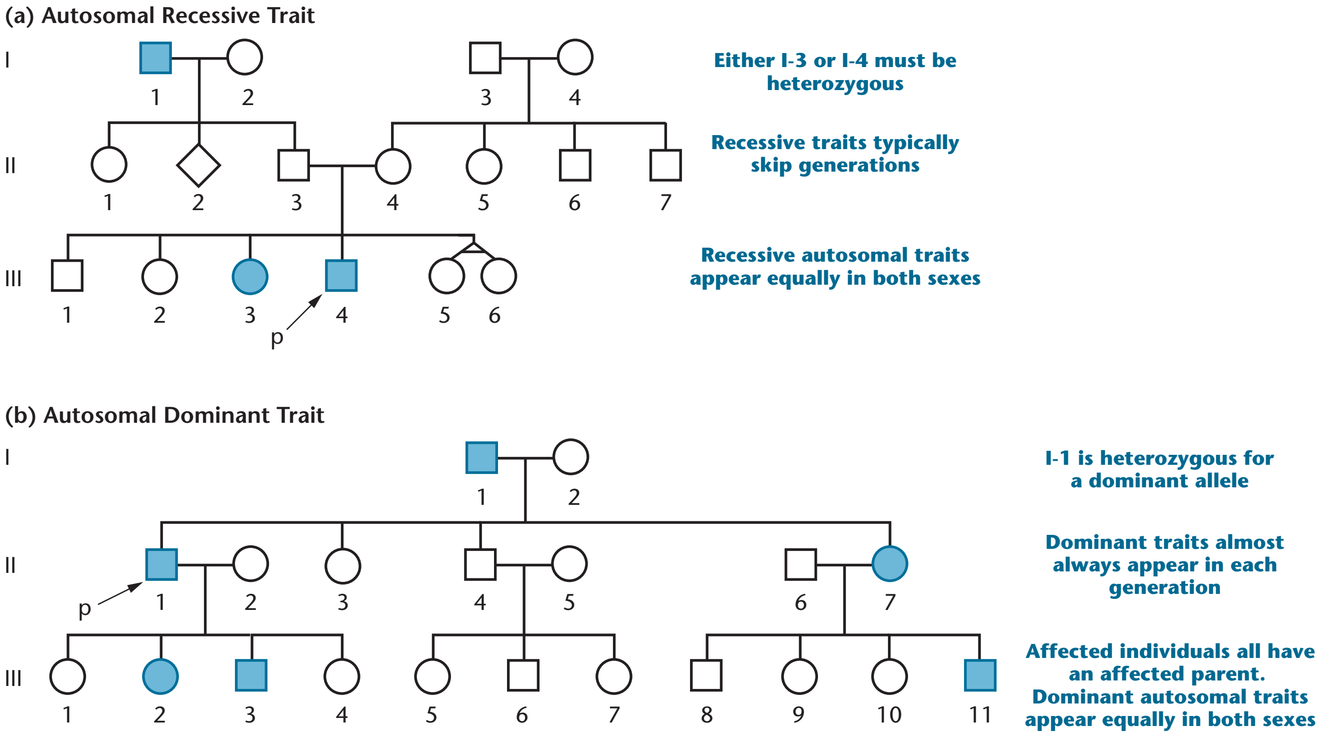
Twins are indicated by diagonal lines stemming from a vertical line connected to the sibship line. For identical, or **monozygotic**, twins, the diagonal lines are linked by a horizontal line. Fraternal, or **dizygotic**, twins lack this connecting line. A number within one of the symbols represents that number of sibs of the same sex and of the same or unknown phenotypes. The individual whose phenotype first brought attention to the family is called the **proband** and is indicated by an arrow connected to the designation **p**. This term applies to either a male or a female.

Pedigree Analysis

In Figure 3.13, two pedigrees are shown. The first is a representative pedigree for a trait that demonstrates autosomal recessive inheritance, such as **albinism**, where synthesis of the pigment melanin in obstructed. The male parent of the first generation (I-1) is affected. Characteristic of a situation in which a parent has a rare recessive trait, the trait “disappears” in the offspring of the next generation. Assuming recessiveness, we might predict that the unaffected female parent (I-2) is a homozygous normal individual because none of the offspring show the disorder. Had she been heterozygous, one-half of the offspring would be expected to exhibit albinism, but none do. However, such a small sample (three offspring) prevents our knowing for certain.

Further evidence supports the prediction of a recessive trait. If albinism were inherited as a dominant trait, individual II-3 would have to express the disorder in order to pass it to his offspring (III-3 and III-4), but he does not. Inspection of the offspring constituting the third generation (row III) provides still further support for the hypothesis that albinism is a recessive trait. If it is, parents II-3 and II-4 are both heterozygous, and approximately one-fourth of their offspring should be affected. Two of the six offspring do show albinism. This deviation from the expected ratio is not unexpected in crosses with few offspring. Once we are confident that albinism is inherited as an autosomal recessive trait, we could portray the II-3 and II-4 individuals with a shaded dot within their larger square and circle. Finally, we can note that, characteristic of pedigrees for autosomal traits, both males and females are affected with equal probability. Later in the text (see Chapter 4), we will examine a pedigree representing a gene located on the sex-determining X chromosome. We will see certain patterns characteristic of the transmission of X-linked traits, such as that these traits are more prevalent in male offspring and are never passed from affected fathers to their sons.

The second pedigree illustrates the pattern of inheritance for a trait such as Huntington disease, which is caused by an autosomal dominant allele. The key to identifying a pedigree that reflects a dominant trait is that all affected offspring will have a parent that also expresses the trait. It is also possible, by chance, that none of the offspring will inherit the dominant allele. If so, the trait will cease to exist


(a) Autosomal Recessive Trait

Either I-3 or I-4 must be heterozygous

Recessive traits typically skip generations

Recessive autosomal traits appear equally in both sexes

(b) Autosomal Dominant Trait

I-1 is heterozygous for a dominant allele

Dominant traits almost always appear in each generation

Affected individuals all have an affected parent. Dominant autosomal traits appear equally in both sexes

FIGURE 3.13 Representative pedigrees for two characteristics, each followed through three generations.

in future generations. Like recessive traits, provided that the gene is autosomal, both males and females are equally affected.

When a given autosomal dominant disease is rare within the population, and most are, then it is highly unlikely that affected individuals will inherit a copy of the mutant gene from both parents. Therefore, in most cases, affected individuals are heterozygous for the dominant allele. As a result, approximately one-half of the offspring inherit it. This is borne out in the second pedigree in Figure 3.13. Furthermore, if a mutation is dominant, and a single copy is sufficient to produce a mutant phenotype, homozygotes are likely to be even more severely affected, perhaps even failing to survive. An illustration of this is the dominant gene for **familial hypercholesterolemia**. Heterozygotes display a defect in their receptors for low-density lipoproteins, the so-called LDLs (known popularly as "bad cholesterol"). As a result, too little cholesterol is taken up by cells from the blood, and elevated plasma levels of LDLs result. Without intervention, such heterozygous individuals usually have heart attacks during the fourth decade of their life, or before. While heterozygotes have LDL levels about double that of a normal individual, rare homozygotes have been detected. They lack LDL receptors altogether, and their LDL levels are nearly ten times above the normal range. They are likely to have a heart attack very early in life, even before age 5, and almost inevitably before they reach the age of 20.

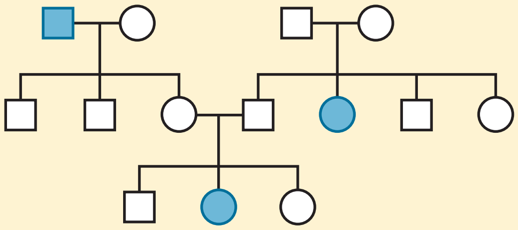
Pedigree analysis of many traits has historically been an extremely valuable research technique in human genetic studies. However, the approach does not usually provide the certainty of the conclusions obtained through experimental crosses yielding large numbers of offspring. Nevertheless, when many independent pedigrees of the same trait or disorder are analyzed, consistent conclusions can often be drawn. **Table 3.2** lists numerous human traits and classifies them according to their recessive or dominant expression.

TABLE 3.2 **Representative Recessive and Dominant Human Traits**

Recessive Traits	Dominant Traits
Albinism	Achondroplasia
Alkaptonuria	Brachydactyly
Color blindness	Ehler–Danlos syndrome
Cystic fibrosis	Hypotrichosis
Duchenne muscular dystrophy	Huntington disease
Galactosemia	Hypercholesterolemia
Hemophilia	Marfan syndrome
Lesch–Nyhan syndrome	Myotonic dystrophy
Phenylketonuria	Neurofibromatosis
Sickle-cell anemia	Phenylthiocarbamide tasting
Tay–Sachs disease	Porphyria (some forms)

NOW SOLVE THIS

3.5 The following pedigree is for myopia (nearsightedness) in humans.



Predict whether the disorder is inherited as the result of a dominant or recessive trait. Determine the most probable genotype for each individual based on your prediction.

■ **HINT:** *This problem asks you to analyze a pedigree and determine the mode of inheritance of myopia. The key to its solution is to identify whether or not there are individuals who express the trait but neither of whose parents also express the trait. Such an observation is a powerful clue and allows you to rule out one mode of inheritance.*

3.10 Mutant Phenotypes Have Been Examined at the Molecular Level

We conclude this chapter by examining two of countless cases where the molecular basis of normal and mutant genes and their resultant phenotypes has now been revealed. Although these explanations were not forthcoming until well over 100 years after Mendel's original work, our discussion of them expands your understanding of how genes control phenotypes. First, we will discuss the molecular basis of one of Mendel's traits, and then we will turn to consideration of a human disorder.

How Mendel's Peas Become Wrinkled: A Molecular Explanation

Only recently, well over a hundred years after Mendel used wrinkled peas in his groundbreaking hybridization experiments, have we come to find out how the *wrinkled* gene makes peas wrinkled. The wild-type allele of the gene encodes a protein called *starch-branching enzyme (SBEI)*. This enzyme catalyzes the formation of highly branched starch molecules as the seed matures.

Wrinkled peas (**Figure 3.14**), which result from the homozygous presence of a mutant form of the gene, lack the activity of this enzyme. As a consequence, the production of branch points is inhibited during the synthesis of starch within the seed, which in turn leads to the accumulation of more sucrose and a higher water content while the seed develops. Osmotic pressure inside the seed rises, causing the seed

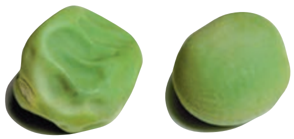


FIGURE 3.14 A wrinkled and round garden pea, the phenotypic traits in one of Mendel's monohybrid crosses.

to lose water, ultimately resulting in a wrinkled appearance at maturity. In contrast, developing seeds that bear at least one copy of the normal gene (being either homozygous or heterozygous for the dominant allele) synthesize starch and achieve an osmotic balance that minimizes the loss of water. The end result for them is a smooth-textured outer coat.

Cloning and analysis of the *SBEI* gene have provided new insight into the relationships between genotypes and phenotypes. Interestingly, the mutant gene contains a foreign sequence of some 800 base pairs that disrupts the normal coding sequence. This foreign segment closely resembles sequences called **transposable elements** that have been discovered to have the ability to move from place to place in the genome of certain organisms. Transposable elements have been found in maize (corn), parsley, snapdragons, and fruit flies, among many other organisms.

Tay—Sachs Disease: The Molecular Basis of a Recessive Disorder in Humans

Of particular interest are cases where a single mutant gene causes multiple effects associated with a severe disorder in humans. Let's consider the modern explanation of the gene that causes **Tay–Sachs disease (TSD)**, a devastating recessive disorder involving unalterable destruction of the central nervous system. Infants with TSD are unaffected at birth and appear to develop normally until they are about 6 months old. Then, a progressive loss of mental and physical abilities occurs. Afflicted infants eventually become blind, deaf, mentally retarded, and paralyzed, often within only a year or two, seldom living beyond age 5. Typical of rare autosomal recessive disorders, two unaffected heterozygous parents, who most often have no family history of the disorder, have a probability of one in four of having a Tay–Sachs child.

We know that proteins are the end products of the expression of most all genes. The protein product involved in TSD has been identified, and we now have a clear understanding of the underlying molecular basis of the disorder. TSD results from the loss of activity of a single enzyme **hexosaminidase A (Hex-A)**. Hex-A, normally found in lysosomes within cells, is needed to break down the ganglioside GM2, a lipid component of nerve cell membranes. Without functional Hex-A, gangliosides accumulate within neurons in the brain and cause deterioration of the nervous system. Heterozygous carriers of TSD with one normal copy of the gene produce only about 50 percent of the normal amount of Hex-A, but they show no symptoms of the disorder. The observation that the activity of only one gene (one wild-type allele) is sufficient for the normal development and function of the nervous system explains and illustrates the molecular basis of recessive mutations. Only when both genes are disrupted by mutation is the mutant phenotype evident. The responsible gene is located on chromosome 15 and codes for the alpha subunit of the Hex-A enzyme. More than 50 different mutations within the gene have been identified that lead to TSD phenotypes.

EXPLORING GENOMICS

Online Mendelian Inheritance in Man

Mastering Genetics Visit the Study Area: Exploring Genomics

The **Online Mendelian Inheritance in Man (OMIM) database** is a catalog of human genes and human disorders that are inherited in a Mendelian manner. Genetic disorders that arise from major chromosomal aberrations, such as monosomy or trisomy (the loss of a chromosome or the presence of a superfluous chromosome, respectively), are not included. The OMIM database, updated daily, is a version of the book *Mendelian Inheritance in Man*, conceived and edited by Dr. Victor McKusick of Johns Hopkins University, until he passed in 2008.

The OMIM entries provide links to a wealth of information, including DNA and protein sequences, chromosomal maps, disease descriptions, and relevant scientific publications. In this exercise, you will explore OMIM to answer questions about the recessive human disease sickle-cell anemia and other Mendelian inherited disorders.

■ Exercise I – Sickle-cell Anemia

In this chapter, you were introduced to recessive and dominant human traits. You will now discover more about sickle-cell anemia as an autosomal

(continued)

Online Mendelian Inheritance in Man—continued

recessive disease by exploring the OMIM database.

1. To begin the search, access the OMIM site at: www.omim.org.
2. In the "Search" box, type "sickle-cell anemia" and click on the "Search" button to perform the search.
3. Click on the link for the entry #603903.
4. Review the text that appears to learn about sickle-cell anemia. Examine the list of subject headings in the left-hand column and explore these links for more information about sickle-cell anemia.
5. Select one or two references at the bottom of the page and follow them to their abstracts in PubMed.
6. Using the information in this entry, answer the following questions:
   a. Which gene is mutated in individuals with sickle-cell anemia?
   b. What are the major symptoms of this disorder?
   c. What was the first published scientific description of sickle-cell anemia?
   d. Describe two other features of this disorder that you learned from the OMIM database, and state where in the database you found this information.

Exercise II – Other Recessive or Dominant Disorders

Select another human disorder that is inherited as either a dominant or recessive trait and investigate its features, following the general steps described in Exercise I. Follow links from OMIM to other databases if you choose.

Describe several interesting pieces of information you acquired during your exploration and cite the information sources you encountered during the search.

CASE STUDY To test or not to test

Thomas discovered a devastating piece of family history when he learned that his brother had been diagnosed with Huntington disease (HD) at age 49. This dominantly inherited autosomal condition usually begins around age 45 with progressive dementia, muscular rigidity, and seizures and ultimately leads to death when affected individuals are in their early 60s. There currently is no effective treatment or cure for this genetic disorder. Thomas, now 38, wonders what the chances are that he also has inherited the mutant allele for HD, leading him to discuss with his wife whether they should seek genetic counseling and whether he should undergo genetic testing. They have two teenage children, a boy and a girl.

1. If they seek genetic counseling, what issues would likely be discussed? Which of these pose grave ethical dilemmas?
2. If you were in Thomas's position, would you want to be tested and possibly learn that you were almost certain to develop the disorder sometime in the next 5–10 years?
3. If Thomas tests positive for the HD allele, should his children be told about the situation, and if so, at what age? Who should make the decision about having the son and daughter tested?

Fulda, K., and Lykens, K. (2006). Ethical Issues in Predictive Genetic Testing: A Public Health Perspective. *J. Med. Ethics* 32:143–147.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Mendel's postulates help describe the basis for the inheritance of phenotypic traits. Based on the analysis of numerous monohybrid crosses, he hypothesized that unit factors exist in pairs and exhibit a dominant/recessive relationship in determining the expression of traits. He further postulated that unit factors segregate during gamete formation, such that each gamete receives one or the other factor, with equal probability.
2. Mendel's postulate of independent assortment, based initially on his analysis of dihybrid crosses, states that each pair of unit factors segregates independently of other such pairs. As a result, all possible combinations of gametes are formed with equal probability.
3. Both the Punnett square and the forked-line method are used to predict the probabilities of phenotypes or genotypes from crosses involving two or more gene pairs. The forked-line method is less complex, but just as accurate as the Punnett square.
4. The discovery of chromosomes in the late 1800s, along with subsequent studies of their behavior during meiosis, led to the rebirth of Mendel's work, linking his unit factors to chromosomes.
5. Since genetic ratios are expressed as probabilities, deriving outcomes of genetic crosses requires an understanding of the laws of probability.
6. Chi-square analysis allows us to assess the null hypothesis, which states that there is no real difference between the expected and observed values. As such, it tests the probability of whether observed variations can be attributed to chance deviation.
7. Pedigree analysis is a method for studying the inheritance pattern of human traits over several generations, providing the basis for predicting the mode of inheritance of characteristics and disorders in the absence of extensive genetic crossing and large numbers of offspring.

INSIGHTS AND SOLUTIONS

As a student, you will be asked to demonstrate your knowledge of transmission genetics by solving various problems. Success at this task requires not only comprehension of theory but also its application to more practical genetic situations. Most students find problem solving in genetics to be both challenging and rewarding. This section is designed to provide basic insights into the reasoning essential to this process.

1. Mendel found that full pea pods are dominant over constricted pods, while round seeds are dominant over wrinkled seeds. One of his crosses was between full, round plants and constricted, wrinkled plants. From this cross, he obtained an F_1 generation that was all full and round. In the F_2 generation, Mendel obtained his classic 9:3:3:1 ratio. Using this information, determine the expected F_1 and F_2 results of a cross between homozygous constricted, round and full, wrinkled plants.

Solution: First, assign gene symbols to each pair of contrasting traits. Use the lowercase first letter of each recessive trait to designate that trait, and use the same letter in uppercase to designate the dominant trait. Thus, *C* and *c* indicate full and constricted pods, respectively, and *W* and *w* indicate the round and wrinkled phenotypes, respectively.

Determine the genotypes of the P_1 generation, form the gametes, combine them in the F_1 generation, and read off the phenotype(s):

$$\begin{array}{llcl}
\textbf{P}_1\textbf{:} & \begin{array}{c}ccWW\\ \text{constricted, round}\\ \downarrow\end{array} & \times & \begin{array}{c}CCww\\ \text{full, wrinkled}\\ \downarrow\end{array}\\
\textbf{Gametes:} & \quad cW & & \quad Cw\\
\textbf{F}_1\textbf{ :} & & \begin{array}{c}CcWw\\ \text{full, round}\end{array} &
\end{array}$$

You can immediately see that the F_1 generation expresses both dominant phenotypes and is heterozygous for both gene pairs. Thus, you expect that the F_2 generation will yield the classic Mendelian ratio of 9:3:3:1. Let's work it out anyway, just to confirm this expectation, using the forked-line method. Both gene pairs are heterozygous and can be expected to assort independently, so we can predict the F_2 outcomes from each gene pair separately and then proceed with the forked-line method.

The F_2 offspring should exhibit the individual traits in the following proportions:

$$\begin{array}{ll}
Cc \times Cc & \qquad Ww \times Ww\\
\downarrow & \qquad \downarrow\\
\left.\begin{array}{l}CC\\ Cc\\ cC\end{array}\right\}\ \text{full} & \qquad \left.\begin{array}{l}WW\\ Ww\\ wW\end{array}\right\}\ \text{round}\\
cc \quad \text{constricted} & \qquad ww \quad \text{wrinkled}
\end{array}$$

Using these proportions to complete a forked-line diagram confirms the 9:3:3:1 phenotypic ratio. (Remember that this ratio represents proportions of 9/16:3/16:3/16:1/16.) Note that we are applying the product law as we compute the final probabilities:

$$\begin{array}{lll}
\text{3/4 full} & \text{3/4 round} & \xrightarrow{(3/4)(3/4)} \text{9/16 full, round}\\
 & \text{1/4 wrinkled} & \xrightarrow{(3/4)(1/4)} \text{3/16 full, wrinkled}\\
\text{1/4 constricted} & \text{3/4 round} & \xrightarrow{(1/4)(3/4)} \text{3/16 constricted, round}\\
 & \text{1/4 wrinkled} & \xrightarrow{(1/4)(1/4)} \text{1/16 constricted, wrinkled}
\end{array}$$

2. In another cross, involving parent plants of unknown genotype and phenotype, the following offspring were obtained.

3/8	full, round
3/8	full, wrinkled
1/8	constricted, round
1/8	constricted, wrinkled

Determine the genotypes and phenotypes of the parents.

Solution: This problem is more difficult and requires keener insight because you must work backward to arrive at the answer. The best approach is to consider the outcomes of pod shape separately from those of seed texture.

Of all the plants, 3/8 + 3/8 = 3/4 are full and 1/8 + 1/8 = 1/4 are constricted. Of the various genotypic combinations that can serve as parents, which will give rise to a ratio of 3/4:1/4? This ratio is identical to Mendel's monohybrid F_2 results, and we can propose that both unknown parents share the same genetic characteristic as the monohybrid F_1 parents: They must both be heterozygous for the genes controlling pod shape, and thus are *Cc*.

Before we accept this hypothesis, let's consider the possible genotypic combinations that control seed texture. If we consider this characteristic alone, we can see that the traits are expressed in a ratio of 3/8 + 1/8 = 1/2 round: 3/8 + 1/8 = 1/2 wrinkled. To generate such a ratio, the parents cannot both be heterozygous or their offspring would yield a 3/4:1/4 phenotypic ratio. They cannot both be homozygous or all offspring would express a single phenotype. Thus, we are left with testing the hypothesis that one parent is homozygous and one is heterozygous for the alleles controlling texture. The potential case of *WW* × *Ww* does not work because it would also yield only a single phenotype. This leaves us with the potential case of *ww* × *Ww*. Offspring in such a mating will yield 1/2 *Ww* (round): 1/2 *ww* (wrinkled), exactly the outcome we are seeking.

Now, let's combine our hypotheses and predict the outcome of the cross. In our solution, we use a dash (–) to indicate that the second allele may be dominant or recessive, since we are only predicting phenotypes.

$$\begin{array}{ll}
3/4\ C– & 1/2\ Ww \rightarrow 3/8\ C–Ww \text{ full, round}\\
 & 1/2\ ww \rightarrow 3/8\ C–ww \text{ full, wrinkled}\\
1/4\ cc & 1/2\ Ww \rightarrow 1/8\ ccWw \text{ constricted, round}\\
 & 1/2\ ww \rightarrow 1/8\ ccww \text{ constricted, wrinkled}
\end{array}$$

(continued)

Insights and Solutions—continued

As you can see, this cross produces offspring in proportions that match our initial information, and we have solved the problem. Note that, in the solution, we have used genotypes in the forked-line method, in contrast to the use of phenotypes in Solution 1.

3. In the laboratory, a genetics student crossed flies with normal long wings with flies expressing the *dumpy* mutation (truncated wings), which she believed was a recessive trait. In the F_1 generation, all flies had long wings. The following results were obtained in the F_2 generation:

792 long-winged flies
208 dumpy-winged flies

The student tested the hypothesis that the dumpy wing is inherited as a recessive trait using χ^2 analysis of the F_2 data.

(a) What ratio was hypothesized?

(b) Did the analysis support the hypothesis?

(c) What do the data suggest about the *dumpy* mutation?

Solution:

(a) The student hypothesized that the F_2 data (792:208) fit Mendel's 3:1 monohybrid ratio for recessive genes.

(b) The initial step in χ^2 analysis is to calculate the expected results (e) for a ratio of 3:1. Then we can compute deviation $o - e$ (d) and the remaining numbers.

Ratio	*o*	*e*	*d*	d^2	d^2/e
3/4	792	750	42	1764	2.35
1/4	208	250	−42	1764	7.06
Total =	1000				

$$\chi^2 = \Sigma \frac{d^2}{e}$$
$$= 2.35 + 7.06$$
$$= 9.41$$

We consult Figure 3.11 to determine the probability (p) and to decide whether the deviations can be attributed to chance. There are two possible outcomes ($n = 2$), so the degrees of freedom (df) $= n - 1$, or 1. The table in Figure 3.11(b) shows that p is a value between 0.01 and 0.001; the graph in Figure 3.11(a) gives an estimate of about 0.001. Since $p < 0.05$, we reject the null hypothesis. The data do not fit a 3:1 ratio.

(c) When the student hypothesized that Mendel's 3:1 ratio was a valid expression of the monohybrid cross, she was tacitly making numerous assumptions. Examining these underlying assumptions may explain why the null hypothesis was rejected. For one thing, she assumed that all the genotypes resulting from the cross were equally viable—that genotypes yielding long wings are equally likely to survive from fertilization through adulthood as the genotype yielding dumpy wings. Further study would reveal that dumpy-winged flies are somewhat less viable than normal flies. As a result, we would expect *less* than 1/4 of the total offspring to express dumpy wings. This observation is borne out in the data, although we have not proven that this is true.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

When working out genetics problems in this and succeeding chapters, always assume that members of the P_1 generation are homozygous, unless the information or data you are given require you to do otherwise.

1. HOW DO WE KNOW? In this chapter, we focused on the Mendelian postulates, probability, and pedigree analysis. We also considered some of the methods and reasoning by which these ideas, concepts, and techniques were developed. On the basis of these discussions, what answers would you propose to the following questions:
(a) How was Mendel able to derive postulates concerning the behavior of "unit factors" during gamete formation, when he could not directly observe them?
(b) How do we know whether an organism expressing a dominant trait is homozygous or heterozygous?
(c) In analyzing genetic data, how do we know whether deviation from the expected ratio is due to chance rather than to another, independent factor?
(d) Since experimental crosses are not performed in humans, how do we know how traits are inherited?

2. CONCEPT QUESTION Review the Chapter Concepts list on p. 36. The first five concepts provide a modern interpretation of Mendelian postulates. Based on these concepts, write a short essay that correlates Mendel's four postulates with what is now known about genes, alleles, and homologous chromosomes.

3. Albinism in humans is inherited as a simple recessive trait. For the following families, determine the genotypes of the parents and offspring. (When two alternative genotypes are possible, list both.)
(a) Two normal parents have five children, four normal and one albino.
(b) A normal male and an albino female have six children, all normal.
(c) A normal male and an albino female have six children, three normal and three albino.
(d) Construct a pedigree of the families in (b) and (c). Assume that one of the normal children in (b) and one of the albino children in (c) become the parents of eight children. Add these children to the pedigree, predicting their phenotypes (normal or albino).

4. Which of Mendel's postulates are illustrated by the pedigree that you constructed in Problem 3? List and define these postulates.

5. Discuss how Mendel's monohybrid results served as the basis for all but one of his postulates. Which postulate was not based on these results? Why?

6. What advantages were provided by Mendel's choice of the garden pea in his experiments?

7. Mendel crossed peas having round seeds and yellow cotyledons (seed leaves) with peas having wrinkled seeds and green cotyledons. All the F_1 plants had round seeds with yellow cotyledons. Diagram this cross through the F_2 generation, using both the Punnett square and forked-line, or branch diagram, methods.

8. Based on the preceding cross, what is the probability that an organism in the F_2 generation will have round seeds and green cotyledons *and* be true breeding?

9. Which of Mendel's postulates can only be demonstrated in crosses involving at least two pairs of traits? State the postulate.

10. In a cross between a black and a white guinea pig, all members of the F_1 generation are black. The F_2 generation is made up of approximately 3/4 black and 1/4 white guinea pigs.
(a) Diagram this cross, showing the genotypes and phenotypes.
(b) What will the offspring be like if two F_2 white guinea pigs are mated?
(c) Two different matings were made between black members of the F_2 generation, with the following results.

Cross	Offspring
Cross 1	All black
Cross 2	3/4 black, 1/4 white

Diagram each of the crosses.

11. What is the basis for homology among chromosomes?

12. In *Drosophila*, *gray* body color is dominant to *ebony* body color, while *long* wings are dominant to *vestigial* wings. Assuming that the P_1 individuals are homozygous, work the following crosses through the F_2 generation, and determine the genotypic and phenotypic ratios for each generation.
(a) gray, long × ebony, vestigial
(b) gray, vestigial × ebony, long
(c) gray, long × gray, vestigial

13. How many different types of gametes can be formed by individuals of the following genotypes: (a) *AaBb*, (b) *AaBB*, (c) *AaBbCc*, (d) *AaBBcc*, (e) *AaBbcc*, and (f) *AaBbCcDdEe?* What are the gametes in each case?

14. Mendel crossed peas having green seeds with peas having yellow seeds. The F_1 generation produced only yellow seeds. In the F_2, the progeny consisted of 6022 plants with yellow seeds and 2001 plants with green seeds. Of the F_2 yellow-seeded plants, 519 were self-fertilized with the following results: 166 bred true for yellow and 353 produced an F_3 ratio of 3/4 yellow: 1/4 green. Explain these results by diagramming the crosses.

15. In a study of black guinea pigs and white guinea pigs, 100 black animals were crossed with 100 white animals, and each cross was carried to an F_2 generation. In 94 of the crosses, all the F_1 offspring were black and an F_2 ratio of 3 black:1 white was obtained. In the other 6 cases, half of the F_1 animals were black and the other half were white. Why? Predict the results of crossing the black and white F_1 guinea pigs from the 6 exceptional cases.

16. Mendel crossed peas having round green seeds with peas having wrinkled yellow seeds. All F_1 plants had seeds that were round and yellow. Predict the results of testcrossing these F_1 plants.

17. Thalassemia is an inherited anemic disorder in humans. Affected individuals exhibit either a minor anemia or a major anemia. Assuming that only a single gene pair and two alleles are involved in the inheritance of these conditions, is thalassemia a dominant or recessive disorder?

18. The following are F_2 results of two of Mendel's monohybrid crosses.

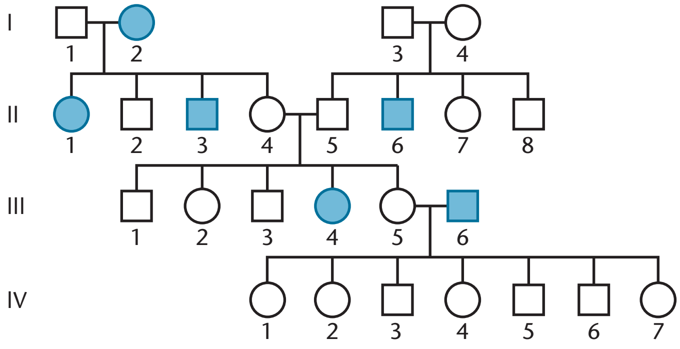
(a) full pods	882
constricted pods	299
(b) violet flowers	705
white flowers	224

For each cross, state a null hypothesis to be tested using χ^2 analysis. Calculate the χ^2 value and determine the *p* value for both. Interpret the *p* values. Can the deviation in each case be attributed to chance or not? Which of the two crosses shows a greater amount of deviation?

19. In assessing data that fell into two phenotypic classes, a geneticist observed values of 250:150. She decided to perform a χ^2 analysis by using the following two different null hypotheses: (a) the data fit a 3:1 ratio, and (b) the data fit a 1:1 ratio. Calculate the χ^2 values for each hypothesis. What can be concluded about each hypothesis?

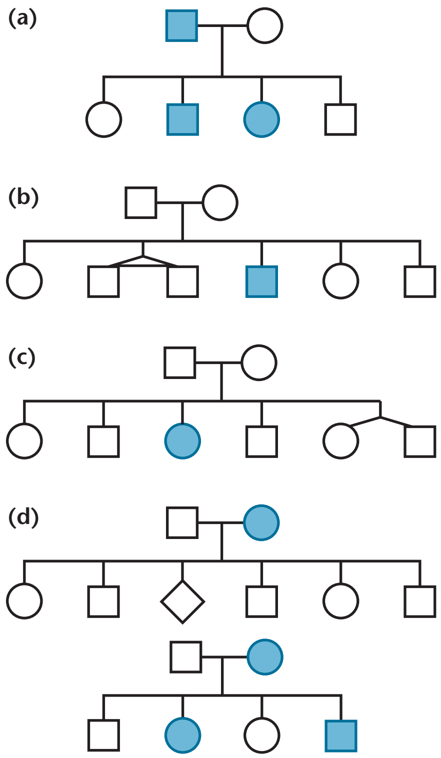
20. The basis for rejecting any null hypothesis is arbitrary. The researcher can set more or less stringent standards by deciding to raise or lower the *p* value used to reject or not reject the hypothesis. In the case of the chi-square analysis of genetic crosses, would the use of a standard of $p = 0.10$ be more or less stringent about not rejecting the null hypothesis? Explain.

21. Consider the following pedigree.



Predict the mode of inheritance of the trait of interest and the most probable genotype of each individual. Assume that the alleles *A* and *a* control the expression.

22. Draw all possible conclusions concerning the mode of inheritance of the trait portrayed in each of the following limited pedigrees. (Each of the four cases is based on a different trait.)



23. For decades scientists have been perplexed by different circumstances surrounding families with rare, early-onset auditory neuropathy (deafness). In some families, parents and grandparents of the proband have normal hearing, while in other families, a number of affected (deaf) family members are scattered throughout the pedigree, appearing in every generation. Assuming a genetic cause for each case, offer a reasonable explanation for the genetic origin of such deafness in the two types of families.

24. A "wrongful birth" case was recently brought before a court in which a child with Smith–Lemli–Opitz syndrome was born to apparently healthy parents. This syndrome is characterized by a cluster of birth defects including cleft palate, and an array of problems with the reproductive and urinary organs. Originally considered by their physician as having a nongenetic basis, the parents decided to have another child, who was also born with Smith–Lemli–Opitz syndrome. In the role of a genetic counselor, instruct the court about what occurred, including the probability of the parents having two affected offspring, knowing that the disorder is inherited as a recessive trait.

Extra-Spicy Problems

25. Tay–Sachs disease (TSD) is an inborn error of metabolism that results in death, often by the age of 2. You are a genetic counselor interviewing a phenotypically normal couple who tell you the male had a female first cousin (on his father's side) who died from TSD and the female had a maternal uncle with TSD. There are no other known cases in either of the families, and none of the matings have been between related individuals. Assume that this trait is very rare.
(a) Draw a pedigree of the families of this couple, showing the relevant individuals.
(b) Calculate the probability that both the male and female are carriers for TSD.
(c) What is the probability that neither of them is a carrier?
(d) What is the probability that one of them is a carrier and the other is not? [*Hint:* The *p* values in (b), (c), and (d) should equal 1.]

26. *Datura stramonium* (the Jimsonweed) expresses flower colors of purple and white and pod textures of smooth and spiny. The results of two crosses in which the parents were not necessarily true breeding are shown below.

white spiny × white spiny	→ 3/4 white spiny: 1/4 white smooth
purple smooth × purple smooth	→ 3/4 purple smooth: 1/4 white smooth

(a) Based on these results, put forward a hypothesis for the inheritance of the purple/white and smooth/spiny traits.
(b) Assuming that true-breeding strains of all combinations of traits are available, what single cross could you execute and carry to an F_2 generation that will prove or disprove your hypothesis? Assuming your hypothesis is correct, what results of this cross will support it?

27. The wild-type (normal) fruit fly, *Drosophila melanogaster,* has straight wings and long bristles. Mutant strains have been isolated that have either curled wings or short bristles. The genes representing these two mutant traits are located on separate chromosomes. Carefully examine the data from the five crosses shown across the top of the next page.
(a) Identify each mutation as either dominant or recessive. In each case, indicate which crosses support your answer.
(b) Assign gene symbols and, for each cross, determine the genotypes of the parents.

	Progeny			
Cross	straight wings, long bristles	straight wings, short bristles	curled wings, long bristles	curled wings, short bristles
1. straight, short × straight, short	30	90	10	30
2. straight, long × straight, long	120	0	40	0
3. curled, long × straight, short	40	40	40	40
4. straight, short × straight, short	40	120	0	0
5. curled, short × straight, short	20	60	20	60

28. To assess Mendel's law of segregation using tomatoes, a true-breeding tall variety (*SS*) is crossed with a true-breeding short variety (*ss*). The heterozygous F_1 tall plants (*Ss*) were crossed to produce two sets of F_2 data, as follows.

Set I	Set II
30 tall	300 tall
5 short	50 short

(a) Using the χ^2 test, analyze the results for both datasets. Calculate χ^2 values and estimate the *p* values in both cases.

(b) From the above analysis, what can you conclude about the importance of generating large datasets in experimental conditions?

29. Albinism, caused by a mutational disruption in melanin (skin pigment) production, has been observed in many species, including humans. In 1991, and again recently in 2017, the only documented observations of an albino humpback whale (named "Migaloo") were observed near New South Wales. Recently, Polanowski and coworkers (Polanowski, A., S. Robinson-Laverick, and D. Paton. (2012). *Journal of Heredity* 103:130–133) studied the genetics of humpback whales from the east coast of Australia, including Migaloo.

(a) Do you think that Migaloo's albinism is more likely caused by a dominant or recessive mutation? Explain your reasoning.

(b) What data would be helpful in determining the answer to part (a)?

30. (a) Assuming that Migaloo's albinism is caused by a rare recessive gene, what would be the likelihood of the establishment of a natural robust subpopulation of albino white humpback whales in this population?

(b) Assuming that Migaloo's albinism is caused by a rare dominant gene, what would be the likelihood of the establishment of a natural robust subpopulation of albino white humpback whales in this population?

4

Labrador retriever puppies expressing brown (chocolate), golden (yellow), and black coat colors, traits controlled by two gene pairs.

Extensions of Mendelian Genetics

CHAPTER CONCEPTS

- While alleles are transmitted from parent to offspring according to Mendelian principles, they often do not display the clear-cut dominant/recessive relationship observed by Mendel.
- In many cases, in a departure from Mendelian genetics, two or more genes are known to influence the phenotype of a single characteristic.
- Still another exception to Mendelian inheritance occurs when genes are located on the X chromosome, because one of the sexes receives only one copy of that chromosome, eliminating the possibility of heterozygosity.
- The result of the various exceptions to Mendelian principles is the occurrence of phenotypic ratios that differ from those produced by standard monohybrid, dihybrid, and trihybrid crosses.
- Phenotypes are often the combined result of genetics and the environment within which genes are expressed.

In Chapter 3, we discussed the fundamental principles of transmission genetics. We saw that genes are present on homologous chromosomes and that these chromosomes segregate from each other and assort independently from other segregating chromosomes during gamete formation. These two postulates are the basic principles of gene transmission from parent to offspring. Once an offspring has received the total set of genes, it is the expression of genes that determines the organism's phenotype. When gene expression does not adhere to a simple dominant/recessive mode, or when more than one pair of genes influences the expression of a single character, the classic 3:1 and 9:3:3:1 F_2 ratios are usually modified. In this chapter, we consider more complex modes of inheritance. In spite of the greater complexity of these situations, the fundamental principles set down by Mendel still hold.

In this chapter, we restrict our initial discussion to the inheritance of traits controlled by only one set of genes. In diploid organisms, which have homologous pairs of chromosomes, two copies of each gene influence such traits. The copies need not be identical since alternative forms of genes, *alleles,* occur within populations. How alleles influence phenotypes will be our primary focus. We will then consider **gene interaction,** a situation in which a single phenotype is affected by more than one set of genes. Numerous examples will be presented to illustrate a variety of heritable patterns observed in such situations.

Thus far, we have restricted our discussion to chromosomes other than the X and Y pair. By examining cases where genes are present on the X chromosome, illustrating **X-linkage,** we will see yet another modification of Mendelian ratios. Our discussion of modified ratios also includes the consideration of sex-limited and sex-influenced inheritance, where the sex of the individual, but not necessarily the genes on the X chromosome, influences the phenotype. We conclude the chapter by showing how a given phenotype often varies depending on the overall environment in which a gene, a cell, or an organism finds itself. This discussion points out that phenotypic expression depends on more than just the genotype of an organism. Please note that some of the topics "discussed" in this chapter are explored in greater depth later in the text (see Chapter 19).

4.1 Alleles Alter Phenotypes in Different Ways

Following the rediscovery of Mendel's work in the early 1900s, research focused on the many ways in which genes influence an individual's phenotype. This course of investigation, stemming from Mendel's findings, is called neo-Mendelian genetics (*neo* from the Greek word meaning "since" or "new").

Each type of inheritance described in this chapter was investigated when observations of genetic data did not conform precisely to the expected Mendelian ratios. Hypotheses that modified and extended the Mendelian principles were proposed and tested with specifically designed crosses. The explanations proffered to account for these observations were constructed in accordance with the principle that a phenotype is under the influence of one or more genes located at specific loci on one or more pairs of homologous chromosomes.

To understand the various modes of inheritance, we must first consider the potential function of an **allele.** An allele is an alternative form of a gene. The allele that occurs most frequently in a population, the one that we arbitrarily designate as normal, is called the *wild-type allele.* This is often, but not always, dominant. Wild-type alleles are responsible for the corresponding wild-type phenotype and are the standards against which all other mutations occurring at a particular locus are compared.

A mutant allele contains modified genetic information and often specifies an altered gene product. For example, in human populations, there are many known alleles of the gene encoding the β chain of human hemoglobin. All such alleles store information necessary for the synthesis of the β chain polypeptide, but each allele specifies a slightly different form of the same molecule. Once the allele's product has been manufactured, the product's function may or may not be altered.

The process of mutation is the source of alleles. For a new allele to be recognized by observation of an organism, the allele must cause a change in the phenotype. A new phenotype results from a change in functional activity of the cellular product specified by that gene. Often, the mutation causes the diminution or the loss of the specific wild-type function. For example, if a gene is responsible for the synthesis of a specific enzyme, a mutation in that gene may ultimately change the conformation of this enzyme and reduce or eliminate its affinity for the substrate. Such a mutation is designated as a **loss-of-function mutation.** If the loss is complete, the mutation has resulted in what is called a **null allele.**

Conversely, other mutations may enhance the function of the wild-type product. Most often when this occurs, it is the result of increasing the quantity of the gene product. For example, the mutation may be affecting the regulation of transcription of the gene under consideration. Such mutations, designated **gain-of-function mutations,** most often result in dominant alleles, since one copy of the mutation in a diploid organism is sufficient to alter the normal phenotype. Examples of gain-of-function mutations include the genetic conversion of *proto-oncogenes,* which regulate the cell cycle, to *oncogenes,* where regulation is overridden by excess gene product. The result is the creation of a cancerous cell. Another example is a mutation that alters the sensitivity of a receptor, whereby an inhibitory signal molecule is unable to quell a particular biochemical response. In a sense, the function of the gene product is always turned on.

Having introduced the concepts of gain- and loss-of-function mutations, we should note the possibility that a mutation will create an allele that produces no detectable change in function. In this case, the mutation would not be immediately apparent since no phenotypic variation would be evident. However, such a mutation could be detected if the DNA sequence of the gene was examined directly. These are sometimes referred to as **neutral mutations** since the gene product presents no change to either the phenotype or the evolutionary fitness of the organism.

Finally, we note that while a phenotypic trait may be affected by a single mutation in one gene, traits are often influenced by many gene products. For example, enzymatic reactions are most often part of complex metabolic pathways leading to the synthesis of an end product, such as an amino acid. Mutations in any of a pathway's reactions can have a common effect—the failure to synthesize the end product. Therefore, phenotypic traits related to the end product are often influenced by more than one gene. Such is the case in *Drosophila* eye color mutations. Eye color results from the synthesis and deposition of a brown and a bright red pigment in the facets of the compound eye. This causes the wild-type eye color to appear brick red. There are a series of recessive loss-of-function mutations that interrupt the multistep pathway leading to the synthesis of the brown pigment. While these mutations represent genes located on different chromosomes, they all result in the same phenotype: a bright red eye whose color is due to the absence of the brown pigment. Examples are the mutations *vermilion, cinnabar,* and *scarlet,* which are indistinguishable phenotypically.

4.2 Geneticists Use a Variety of Symbols for Alleles

Earlier in the text, we learned a standard convention used to symbolize alleles for very simple Mendelian traits (see Chapter 3). The initial letter of the name of a recessive trait, lowercased and italicized, denotes the recessive allele, and the same letter in uppercase refers to the dominant allele. Thus, in the case of tall and dwarf, where dwarf is recessive, *D* and *d* represent the alleles responsible for these respective traits. Mendel used upper- and lowercase letters such as these to symbolize his unit factors.

Another useful system was developed in genetic studies of the fruit fly *Drosophila melanogaster* to discriminate between wild-type and mutant traits. This system uses the initial letter, or a combination of several letters, from the name of the mutant trait. If the trait is recessive, lowercase is used; if it is dominant, uppercase is used. The contrasting wild-type trait is denoted by the same letters, but with a superscript +. For example, *ebony* is a recessive body color mutation in *Drosophila.* The normal wild-type body color is gray. Using this system, we denote *ebony* by the symbol *e,* while gray is denoted by e^+. The responsible locus may be occupied by either the wild-type allele (e^+) or the mutant allele (*e*). A diploid fly may thus exhibit one of three possible genotypes (the two phenotypes are indicated parenthetically):

e^+/e^+	**gray homozygote (wild type)**
e^+/e	**gray heterozygote (wild type)**
$e\ /e$	**ebony homozygote (mutant)**

The slash between the letters indicates that the two allele designations represent the same locus on two homologous chromosomes. If we instead consider a mutant allele that is dominant to the normal wild-type allele, such as *Wrinkled* wing in *Drosophila,* the three possible genotypes are *Wr*/*Wr*, Wr/Wr^+, and Wr^+/Wr^+. The initial two genotypes express the mutant wrinkled-wing phenotype.

One advantage of this system is that further abbreviation can be used when convenient: The wild-type allele may simply be denoted by the + symbol. With *ebony* as an example, the designations of the three possible genotypes become the following:

$+/+$	**gray homozygote (wild type)**
$+/e$	**gray heterozygote (wild type)**
$e\ /e$	**ebony homozygote (mutant)**

Another variation is utilized when no dominance exists between alleles (a situation we will explore in Section 4.3). We simply use uppercase letters and superscripts to denote alternative alleles (e.g., R^1 and R^2, L^M and L^N, and I^A and I^B).

Many diverse systems of genetic nomenclature are used to identify genes in various organisms. Usually, the symbol selected reflects the function of the gene or even a disorder caused by a mutant gene. For example, in yeast, *cdk* is the abbreviation for the *c*yclin-*d*ependent *k*inase gene, whose product is involved in the cell-cycle regulation mechanism (discussed in Chapter 2). In bacteria, leu^- refers to a mutation that interrupts the biosynthesis of the amino acid leucine, and the wild-type gene is designated leu^+. The symbol *dnaA* represents a bacterial gene involved in DNA replication (and DnaA, without italics, is the protein made by that gene). In humans, italicized capital letters are used to name genes: *BRCA1* represents one of the genes associated with susceptibility to *br*east *ca*ncer. Although these different systems may seem complex, they are useful ways to symbolize genes.

4.3 Neither Allele Is Dominant in Incomplete, or Partial, Dominance

Unlike the Mendelian crosses (reported in Chapter 3), a cross between parents with contrasting traits may sometimes generate offspring with an intermediate phenotype. For example, if a four-o'clock or a snapdragon plant with red flowers is crossed with a white-flowered plant, the offspring have pink flowers. Because some red pigment is produced in the F_1 intermediate-colored plant, neither the red nor white flower color is dominant. Such a situation is known as **incomplete,** or **partial, dominance.**

If the phenotype is under the control of a single gene and two alleles, where neither is dominant, the results of the F_1 (pink) × F_1 (pink) cross can be predicted. The resulting F_2 generation shown in **Figure 4.1** confirms the hypothesis that only one pair of alleles determines these phenotypes. The genotypic ratio (1:2:1) of the F_2 generation is identical to that of Mendel's monohybrid cross. However, because neither allele is dominant, the phenotypic ratio is identical to the genotypic ratio (in contrast to the 3:1 phenotypic ratio of a Mendelian monohybrid cross). Note that because neither allele is recessive, we have chosen not to use upper- and lowercase letters as symbols. Instead, we denote the alleles responsible for red and white color as R^1 and R^2. We could have chosen W^1 and W^2 or still other designations such as C^W and C^R, where *C* indicates "color" and the *W* and *R* superscripts indicate "white" and "red," respectively.

How are we to interpret lack of dominance whereby an intermediate phenotype characterizes heterozygotes? The most accurate way is to consider gene expression in a quantitative way. In the case of flower color, the mutation causing

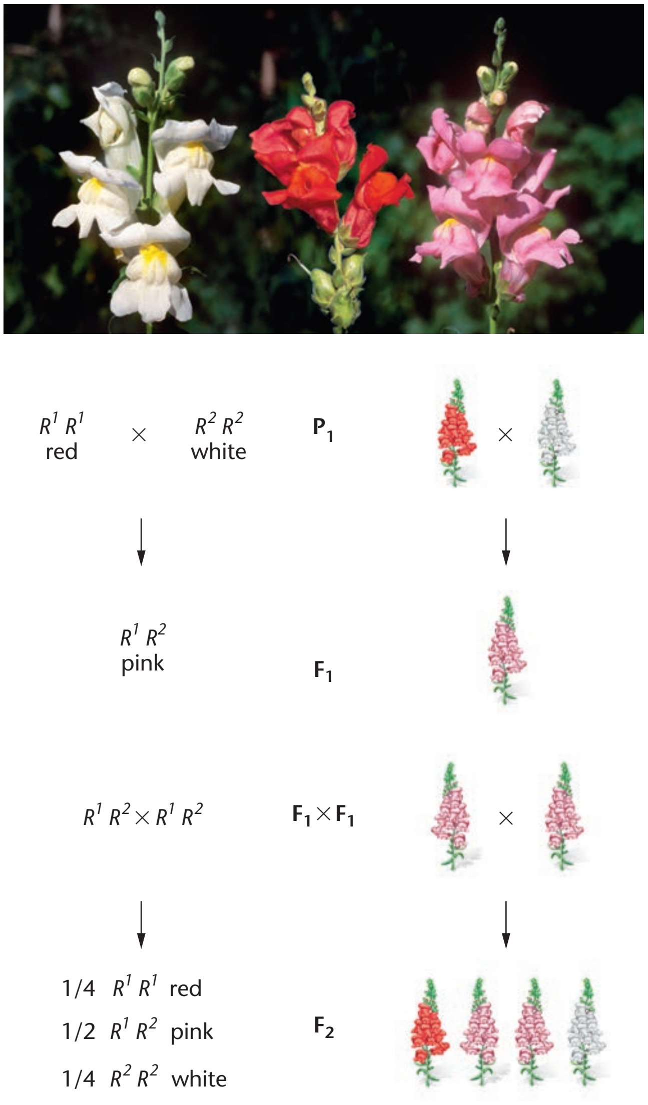


FIGURE 4.1 Incomplete dominance shown in the flower color of snapdragons.

white flowers is most likely one where complete "loss of function" occurs. In this case, it is likely that the gene product of the wild-type allele (R^1) is an enzyme that participates in a reaction leading to the synthesis of a red pigment. The mutant allele (R^2) produces an enzyme that cannot catalyze the reaction leading to pigment. The end result is that the heterozygote produces only about half the pigment of the red-flowered plant and the phenotype is pink.

Clear-cut cases of incomplete dominance are relatively rare. However, even when one allele seems to have complete dominance over the other, careful examination of the gene product and its activity, rather than the phenotype, often reveals an intermediate level of gene expression. An example is the human biochemical disorder **Tay–Sachs disease,** previously discussed in Chapter 3 (see p. 55), in which homozygous recessive individuals are severely affected with a fatal lipid-storage disorder and neonates die during their first one to three years of life. In afflicted individuals, there is almost no activity of **hexosaminidase A,** an enzyme normally involved in lipid metabolism. Heterozygotes, with only a single copy of the mutant gene, are phenotypically normal, but with only about 50 percent of the enzyme activity found in homozygous normal individuals. Fortunately, this level of enzyme activity is adequate to achieve normal biochemical function. This situation is not uncommon in enzyme disorders and illustrates the concept of the **threshold effect,** whereby normal phenotypic expression occurs anytime a minimal level of gene product is attained. Most often, and in particular in Tay–Sachs disease, the threshold is less than 50 percent.

4.4 In Codominance, the Influence of Both Alleles in a Heterozygote Is Clearly Evident

If two alleles of a single gene are responsible for producing two distinct, detectable gene products, a situation different from incomplete dominance or dominance/recessiveness arises. In this case, *the joint expression of both alleles in a heterozygote* is called **codominance.** The **MN blood group** in humans illustrates this phenomenon. Karl Landsteiner and Philip Levine discovered a glycoprotein molecule found on the surface of red blood cells that acts as a native antigen, providing biochemical and immunological identity to individuals. In the human population, two forms of this glycoprotein exist, designated M and N; an individual may exhibit either one or both of them.

The MN system is under the control of a locus found on chromosome 4, with two alleles designated L^M and L^N. Because humans are diploid, three combinations are possible, each resulting in a distinct blood type:

Genotype	Phenotype
$L^M L^M$	M
$L^M L^N$	MN
$L^N L^N$	N

As predicted, a mating between two heterozygous MN parents may produce children of all three blood types, as follows:

$$
\begin{gathered}
L^M L^N \times L^M L^N \\
\downarrow \\
1/4\, L^M L^M \\
1/2\, L^M L^N \\
1/4\, L^N L^N
\end{gathered}
$$

Once again, the genotypic ratio 1:2:1 is upheld.

Codominant inheritance is characterized by *distinct expression of the gene products of both alleles.* This characteristic distinguishes codominance from incomplete dominance, where heterozygotes express an intermediate, blended phenotype. For codominance to be studied, both products must be phenotypically detectable. We shall see another example of codominance when we examine the ABO blood-type system.

4.5 Multiple Alleles of a Gene May Exist in a Population

The information stored in any gene is extensive, and mutations can modify this information in many ways. Each change produces a different allele. Therefore, for any gene, the number of alleles within members of a population need not be restricted to two. When three or more alleles of the same gene—which we designate as **multiple alleles**—are present in a population, the resulting mode of inheritance may be unique. It is important to realize that *multiple alleles can be studied only in populations.* Any individual diploid organism has, at most, two homologous gene loci that may be occupied by different alleles of the same gene. However, among members of a species, numerous alternative forms of the same gene can exist.

The ABO Blood Groups

The simplest case of multiple alleles occurs when three alternative alleles of one gene exist. This situation is illustrated in the inheritance of the **ABO blood groups** in humans, discovered by Karl Landsteiner in the early 1900s. The ABO system, like the MN blood types, is characterized by the presence of antigens on the surface of red blood cells. The A and B antigens are distinct from the MN antigens and are under the control of a different gene, located on chromosome 9. As in the MN system, one combination of alleles in the ABO system exhibits a codominant mode of inheritance.

The ABO phenotype of any individual is ascertained by mixing a blood sample with an antiserum containing type A or type B antibodies. If an antigen is present on the surface of the person's red blood cells, it will react with the corresponding antibody and cause clumping, or agglutination, of the red blood cells. When an individual is tested in this way, one of four phenotypes may be revealed. Each individual has the A antigen (A phenotype), the B antigen (B phenotype), the A and B antigens (AB phenotype), or neither antigen (O phenotype).

In 1924, it was hypothesized that these phenotypes were inherited as the result of three alleles of a single gene. This hypothesis was based on studies of the blood types of many different families. Although different designations can be used, we will use the symbols I^A, I^B, and *i* to distinguish these three alleles. The *I* designation stands for *isoagglutinogen,* another term for antigen. If we assume that the I^A and I^B alleles are responsible for the production of their respective A and B antigens and that *i* is an allele that does not produce any detectable A or B antigens, we can list the various genotypic possibilities and assign the appropriate phenotype to each:

Genotype	Antigen	Phenotype
$I^A I^A$	A	A
$I^A i$	A	
$I^B I^B$	B	B
$I^B i$	B	
$I^A I^B$	A, B	AB
$i\,i$	Neither	O

In these assignments, the I^A and I^B alleles are dominant to the *i* allele, but codominant to each other.

Our knowledge of human blood types has several practical applications, including compatible blood transfusions and successful organ transplants.

The A and B Antigens

The biochemical basis of the ABO blood type system has now been carefully worked out. The A and B antigens are actually carbohydrate groups (sugars) that are bound to lipid molecules (fatty acids) protruding from the membrane of the red blood cell. The specificity of the A and B antigens is based on the terminal sugar of the carbohydrate group.

Almost all individuals possess what is called the **H substance,** to which one or two terminal sugars are added. As shown in **Figure 4.2**, the H substance itself contains three sugar molecules—galactose (Gal), *N*-acetylglucosamine (AcGluNH), and fucose—chemically linked together. The I^A allele is responsible for an enzyme that can add the terminal sugar *N*-acetylgalactosamine (AcGalNH) to the H substance. The I^B allele is responsible for a modified enzyme that cannot add *N*-acetylgalactosamine, but instead can add a terminal galactose. Heterozygotes (I^AI^B) add either one or the other sugar at the many sites (substrates) available on the surface of the red blood cell, illustrating the biochemical basis of codominance in individuals of the AB blood type. Finally, persons of type O (*ii*) cannot add either terminal sugar; these persons have only the H substance protruding from the surface of their red blood cells.

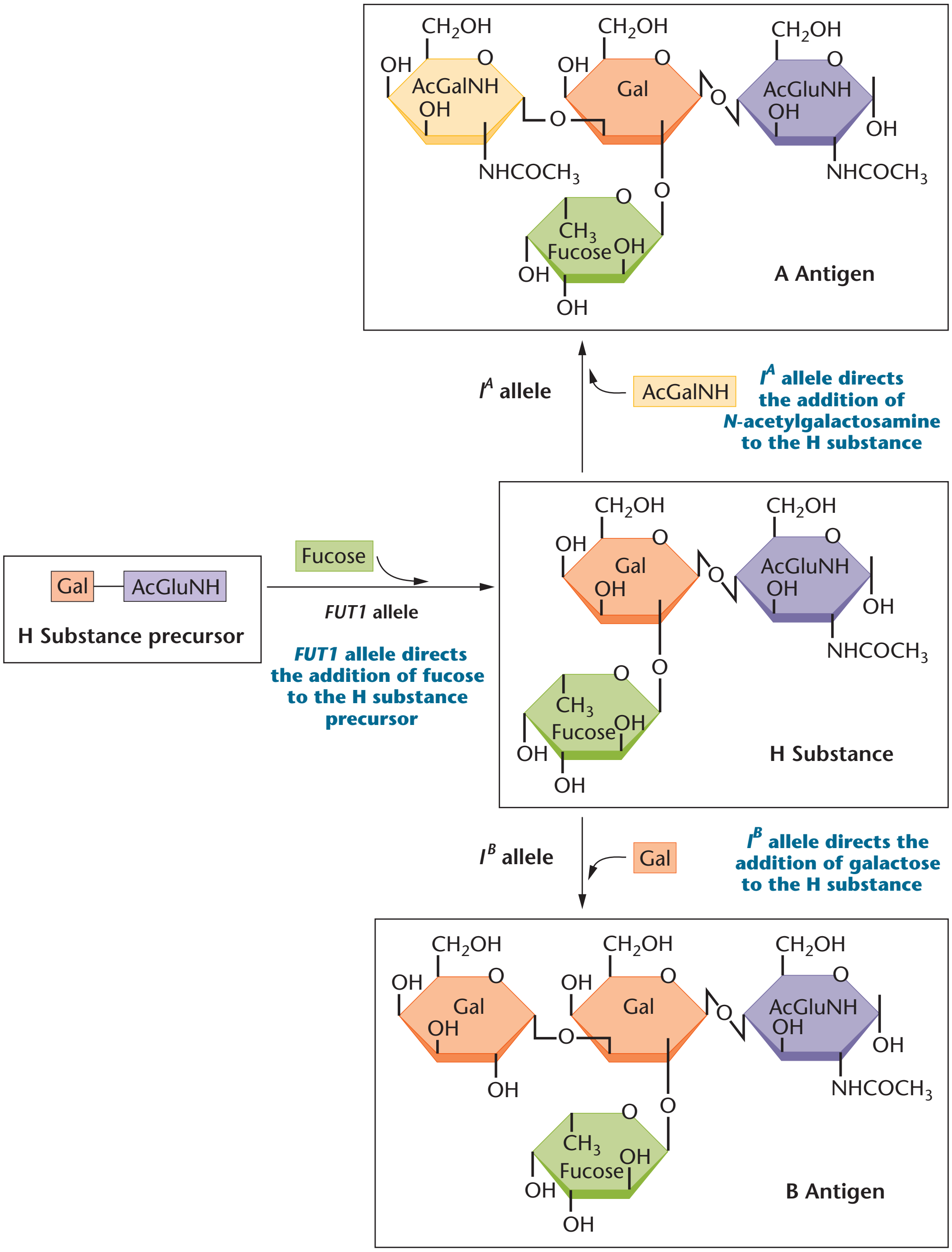

CH$_2$OH
O
OH
AcGalNH
OH
NHCOCH$_3$
Gal
AcGluNH
CH$_3$
Fucose
A Antigen
I^A allele
AcGalNH
I^A allele directs the addition of *N*-acetylgalactosamine to the H substance
Gal
AcGluNH
H Substance precursor
Fucose
FUT1 allele
FUT1 allele directs the addition of fucose to the H substance precursor
H Substance
I^B allele
Gal
I^B allele directs the addition of galactose to the H substance
B Antigen

FIGURE 4.2 The biochemical basis of the ABO blood groups. The wild-type *FUT1* allele, present in almost all humans, directs the conversion of a precursor molecule to the H substance by adding a molecule of fucose to it. The I^A and I^B alleles are then able to direct the addition of terminal sugar residues to the H substance. The *i* allele is unable to direct either of these terminal additions. Failure to produce the H substance results in the Bombay phenotype, in which individuals are type O regardless of the presence of an I^A or I^B allele. Gal: galactose; AcGluNH: *N*-acetylglucosamine; AcGalNH: *N*-acetylgalactosamine.

The Bombay Phenotype

In 1952, a very unusual situation provided information concerning the genetic basis of the H substance. A woman in Bombay displayed a unique genetic history inconsistent with her blood type. In need of a transfusion, she was found to lack both the A and B antigens and was thus typed as O. However, as shown in the partial pedigree in **Figure 4.3**, one of her parents was type AB, and she herself was the obvious donor of an I^B allele to two of her offspring. Thus, she was genetically type B but functionally type O!

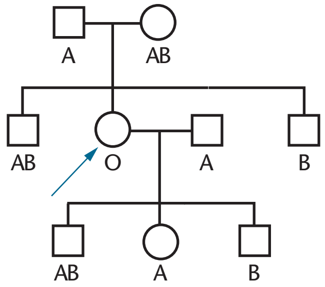


FIGURE 4.3 A partial pedigree of a woman with the Bombay phenotype. Functionally, her ABO blood group behaves as type O. Genetically, she is type B.

This woman was subsequently shown to be homozygous for a rare recessive mutation in a gene designated *FUT1* (encoding an enzyme, fucosyl transferase), which prevented her from synthesizing the complete H substance. In this mutation, the terminal portion of the carbohydrate chain protruding from the red cell membrane lacks fucose, normally added by the enzyme. In the absence of fucose, the enzymes specified by the I^A and I^B alleles apparently are unable to recognize the incomplete H substance as a proper substrate. Thus, neither the terminal galactose nor *N*-acetylgalactosamine can be added, even though the appropriate enzymes capable of doing so are present and functional. As a result, the ABO system genotype cannot be expressed in individuals homozygous for the mutant form of the *FUT1* gene; even though they may have the I^A and/or the I^B alleles, neither antigen is added to the cell surface, and they are functionally type O. To distinguish them from the rest of the population, they are said to demonstrate the **Bombay phenotype.** The frequency of the mutant *FUT1* allele is exceedingly low. Hence, the vast majority of the human population can synthesize the H substance.

The *white* Locus in *Drosophila*

Many other phenotypes in plants and animals are influenced by multiple allelic inheritance. In *Drosophila,* many alleles are present at practically every locus. The recessive mutation that causes white eyes, discovered by Thomas H. Morgan and Calvin Bridges in 1912, is one of over 100 alleles that can occupy this locus. In this allelic series, eye colors range from complete absence of pigment in the *white* allele to deep ruby in the *white-satsuma* allele, orange in the *white-apricot* allele, and a buff color in the *white-buff* allele. These alleles are designated w, w^{sat}, w^a, and w^{bf}, respectively. In each case, the total amount of pigment in these mutant eyes is reduced to less than 20 percent of that found in the brick-red wild-type eye. **Table 4.1** lists these and other *white* alleles and their color phenotypes.

TABLE 4.1 Some of the Alleles Present at the *white* Locus of *Drosophila*

Allele	Name	Eye Color
w	*white*	pure white
w^a	*white-apricot*	yellowish orange
w^{bf}	*white-buff*	light buff
w^{bl}	*white-blood*	yellowish ruby
w^{cf}	*white-coffee*	deep ruby
w^e	*white-eosin*	yellowish pink
w^{mo}	*white-mottled orange*	light mottled orange
w^{sat}	*white-satsuma*	deep ruby
w^{sp}	*white-spotted*	fine grain, yellow mottling
w^t	*white-tinged*	light pink

It is interesting to note the biological basis of the original *white* mutation in *Drosophila*. Given what we know about eye color in this organism, it might be logical to presume that the mutant allele somehow interrupts the biochemical synthesis of pigments making up the brick-red eye of the wild-type fly. However, it is now clear that the product of the *white* locus is a protein that is involved in transporting pigments into the ommatidia (the individual units) comprising the compound eye. While flies expressing the *white* mutation can synthesize eye pigments normally, they cannot transport them into these structural units of the eye, thus rendering the white phenotype.

NOW SOLVE THIS

4.1 In the guinea pig, one locus involved in the control of coat color may be occupied by any of four alleles: C (full color), c^k (sepia), c^d (cream), or c^a (albino), with an order of dominance of: $C > c^k > c^d > c^a$. (C is dominant to all others, c^k is dominant to c^d and c^a, but not C, etc.) In the following crosses, determine the parental genotypes and predict the phenotypic ratios that would result:

(a) sepia × cream, where both guinea pigs had an albino parent

(b) sepia × cream, where the sepia guinea pig had an albino parent and the cream guinea pig had two sepia parents

(c) sepia × cream, where the sepia guinea pig had two full-color parents and the cream guinea pig had two sepia parents

(d) sepia × cream, where the sepia guinea pig had a full-color parent and an albino parent and the cream guinea pig had two full-color parents

■ **HINT:** *This problem involves an understanding of multiple alleles. The key to its solution is to note particularly the hierarchy of dominance of the various alleles. Remember also that even though there can be more than two alleles in a population, an individual can have at most two of these. Thus, the allelic distribution into gametes adheres to the principle of segregation.*

4.6 Lethal Alleles Represent Essential Genes

Many gene products are essential to an organism's survival. Mutations resulting in the synthesis of a gene product that is nonfunctional can often be tolerated in the heterozygous state; that is, one wild-type allele may be sufficient to produce enough of the essential product to allow survival. However, such a mutation behaves as a **recessive lethal allele,** and homozygous recessive individuals will not survive. The time of death will depend on when the product is essential. In mammals, for example, this might occur during development, early childhood, or even adulthood.

In some cases, the allele responsible for a lethal effect when homozygous may also result in a distinctive mutant phenotype when present heterozygously. *It is behaving as a recessive lethal allele but is dominant with respect to the phenotype.* For example, a mutation that causes a yellow coat in mice was discovered in the early part of this century. The yellow coat varies from the normal agouti (wild-type) coat phenotype, as shown in **Figure 4.4**. Crosses between the various combinations of the two strains yield unusual results:

		Crosses		
(A) agouti	×	agouti	⟶	all agouti
(B) yellow	×	yellow	⟶	2/3 yellow 1/3 agouti
(C) agouti	×	yellow	⟶	1/2 yellow 1/2 agouti

These results are explained on the basis of a single pair of alleles. With regard to coat color, the mutant *yellow* allele (A^Y) is dominant to the wild-type *agouti* allele (A), so heterozygous mice will have yellow coats. However, the *yellow* allele is also a homozygous recessive lethal. When present

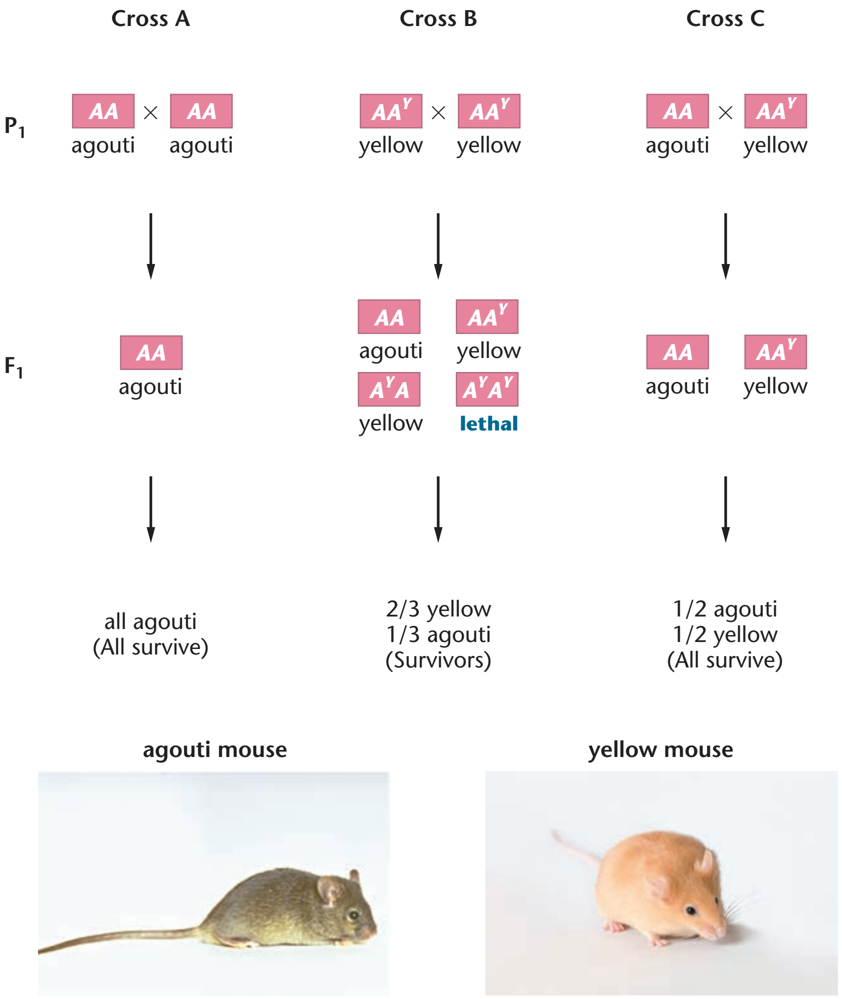


FIGURE 4.4 Inheritance patterns in three crosses involving the *agouti* allele (A) and the mutant *yellow* allele (A^Y) in the mouse. Note that the mutant allele behaves dominantly to the normal allele in controlling coat color, but it also behaves as a homozygous recessive lethal allele. Mice with the genotype A^YA^Y do not survive.

in two copies, the mice die before birth. Thus, there are no homozygous yellow mice. The genetic basis for these three crosses is shown in Figure 4.4.

In other cases, a mutation may behave as a **dominant lethal allele.** In such cases, the presence of just one copy of the allele results in the death of the individual. In humans, a disorder called **Huntington disease** is due to a dominant autosomal allele *H,* where the onset of the disease in heterozygotes (*Hh*) is delayed, usually well into adulthood. Affected individuals then undergo gradual nervous and motor degeneration until they die. This lethal disorder is particularly tragic because it has such a late onset, typically at about age 40. By that time, the affected individual may have produced a family, and each of their children has a 50 percent probability of inheriting the lethal allele, transmitting the allele to his or her offspring, and eventually developing the disorder. The American folk singer and composer Woody Guthrie (father of folk singer Arlo Guthrie) died from this disease at age 55.

Dominant lethal alleles are rarely observed. For these alleles to exist in a population, the affected individuals must reproduce before the lethal allele is expressed, as can occur in Huntington disease. If all affected individuals die before reaching reproductive age, the mutant gene will not be passed to future generations, and the mutation will disappear from the population unless it arises again as a result of a new mutation.

The Molecular Basis of Dominance, Recessiveness, and Lethality: The *agouti* Gene

Molecular analysis of the gene resulting in the agouti and yellow mice has provided insight into how a mutation can be both dominant for one phenotypic effect (hair color) and recessive for another (embryonic development). The A^Y allele is a classic example of a gain-of-function mutation. Animals homozygous for the wild-type *A* allele have yellow pigment deposited as a band on the otherwise black hair shaft, resulting in the agouti phenotype (see Figure 4.4). Heterozygotes deposit yellow pigment along the entire length of hair shafts as a result of the deletion of the regulatory region preceding the DNA coding region of the A^Y allele. Without any means to regulate its expression, one copy of the A^Y allele is always turned on in heterozygotes, resulting in the gain of function leading to the dominant effect.

The homozygous lethal effect has also been explained by molecular analysis of the mutant gene. The extensive deletion of genetic material that produced the A^Y allele actually extends into the coding region of an adjacent gene (*Merc*), rendering it nonfunctional. It is this gene that is critical to embryonic development, and the loss of its function in A^Y/A^Y homozygotes is what causes lethality. Heterozygotes exceed the threshold level of the wild-type Merc gene product and thus survive.

4.7 Combinations of Two Gene Pairs with Two Modes of Inheritance Modify the 9:3:3:1 Ratio

Each example discussed so far modifies Mendel's 3:1 F_2 monohybrid ratio. Therefore, combining any two of these modes of inheritance in a dihybrid cross will also modify the classical 9:3:3:1 dihybrid ratio. Having established the foundation of the modes of inheritance of incomplete dominance, codominance, multiple alleles, and lethal alleles, we can now deal with the situation of two modes of inheritance occurring simultaneously. Mendel's principle of independent assortment applies to these situations, provided that the genes controlling each character are not located on the same chromosome—in other words, that they do not demonstrate what is called *genetic linkage.*

Consider, for example, a mating between two humans who are both heterozygous for the autosomal recessive gene that causes albinism and who are both of blood type AB. What is the probability of a particular phenotypic combination occurring in each of their children? Albinism is inherited in the simple Mendelian fashion, and the blood types are determined by the series of three multiple alleles, I^A, I^B, and *i*. The solution to this problem is diagrammed in **Figure 4.5** using the forked-line method. This dihybrid cross does not yield four phenotypes in the classical 9:3:3:1 ratio. Instead, six phenotypes occur in a 3:6:3:1:2:1 ratio, establishing the expected probability for each phenotype. This is just one of the many variants of modified ratios that are possible when different modes of inheritance are combined.

EVOLVING CONCEPT OF A GENE

Based on the work of many geneticists following the rediscovery of Mendel's work in the very early part of the twentieth century, the chromosome theory of inheritance was put forward, which hypothesized that chromosomes are the carriers of genes and that meiosis is the physical basis of Mendel's postulates. In the ensuing 40 years, the concept of a gene evolved to reflect that this hereditary unit can exist in multiple forms, or alleles, each of which can impact on the phenotype in different ways, leading to incomplete dominance, codominance, and even lethality. It became clear that the process of mutation was the source of new alleles. ■

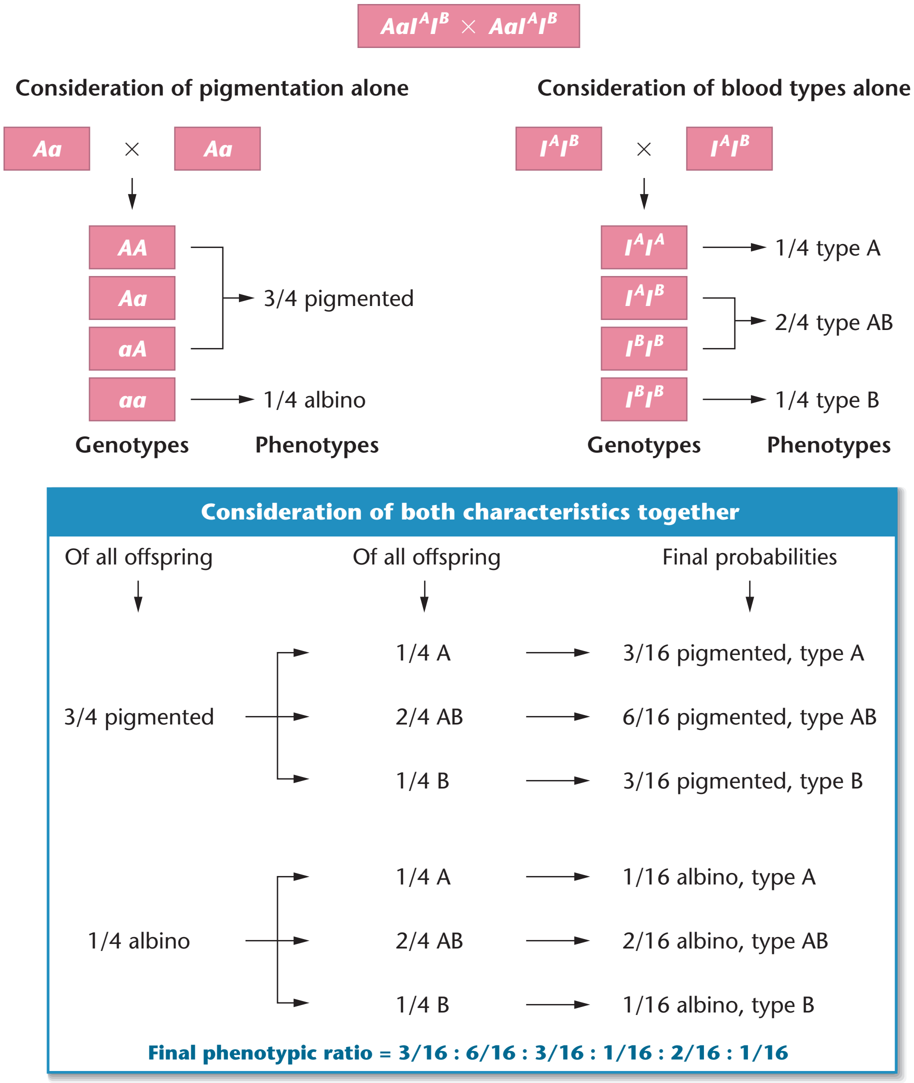


FIGURE 4.5 Calculation of the probabilities in a mating involving the ABO blood type and albinism in humans, using the forked-line method.

4.8 Phenotypes Are Often Affected by More Than One Gene

Soon after Mendel's work was rediscovered, experimentation revealed that in many cases a given phenotype is affected by more than one gene. This was a significant discovery because it revealed that genetic influence on the phenotype is often much more complex than the situations Mendel encountered in his crosses with the garden pea. Instead of single genes controlling the development of individual parts of a plant or animal body, it soon became clear that phenotypic characters such as eye color, hair color, or fruit shape can be influenced by many different genes and their products.

The term **gene interaction** is often used to express the idea that several genes influence a particular characteristic. This does not mean, however, that two or more genes or their products necessarily interact directly with one another to influence a particular phenotype. Rather, the term means that the cellular function of numerous gene products contributes to the development of a common phenotype. For example, the development of an organ such as the eye of an insect is exceedingly complex and leads to a structure with multiple phenotypic manifestations, for example, to an eye having a specific size, shape, texture, and color. The development of the eye is a complex cascade of developmental events leading to that organ's formation. This process illustrates the developmental concept of **epigenesis,** whereby each step of development increases the complexity of the organ or feature of interest and is under the control and influence of many genes.

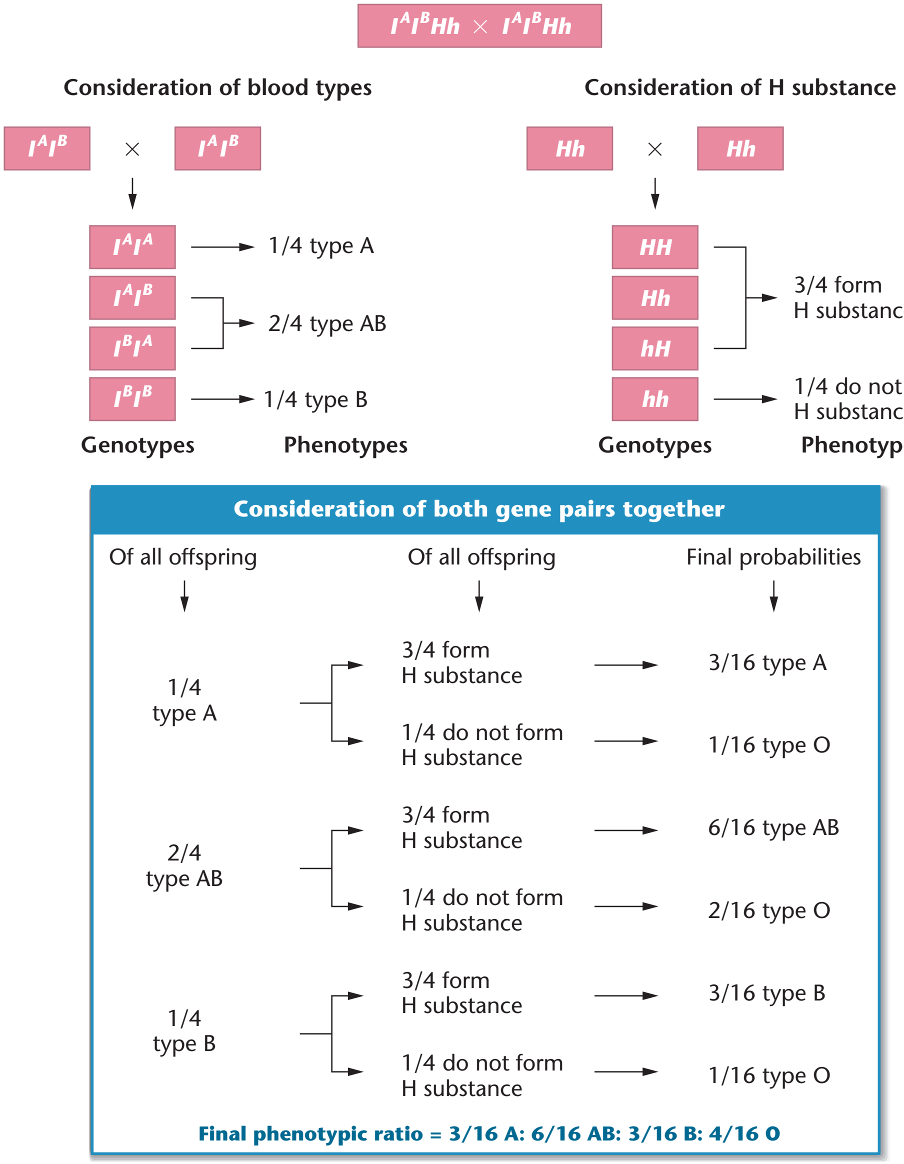
An enlightening example of epigenesis and multiple gene interaction involves the formation of the inner ear in mammals, allowing organisms to detect and interpret sound. The structure and function of the inner ear are exceedingly complex. Its formation includes distinctive anatomical features not only to capture, funnel, and transmit external sound toward and through the middle ear, but also to convert sound waves into nerve impulses within the inner ear. Thus, the ear forms as a result of a cascade of intricate developmental events influenced by many genes. Mutations that interrupt many of the steps of ear development lead to a common phenotype: **hereditary deafness.** In a sense, these many genes "interact" to produce a common phenotype. In such situations, the mutant phenotype is described as a **heterogeneous trait,** reflecting the many genes involved. In humans, while a few common alleles are responsible for the vast majority of cases of hereditary deafness, over 50 genes are involved in the development of the ability to discern sound.

Epistasis

Some of the best examples of gene interaction are those showing the phenomenon of **epistasis,** where the expression of one gene masks or modifies the effect of a second gene. Sometimes the genes involved influence the same general phenotypic characteristic in an antagonistic manner, which leads to masking. In other cases, however, the genes involved exert their influence on one another in a complementary, or cooperative, fashion.

For example, the homozygous presence of a recessive allele may prevent or override the expression of other alleles at a second locus (or several other loci). In this case, the alleles at the first locus are said to be *epistatic* to those at the second locus, and the alleles at the second locus are *hypostatic* to those at the first locus. As we will see, there are several variations on this theme. In another example, a single dominant allele at the first locus may be epistatic to the expression of the alleles at a second gene locus. In a third example, two genes may *complement* one another such that at least one dominant allele in each gene is required to express a particular phenotype.

The Bombay phenotype discussed earlier is an example of the homozygous recessive condition at one locus masking the expression of a second locus. There we established that the homozygous presence of the mutant form of the *FUT1* gene masks the expression of the I^A and I^B alleles. Only individuals containing at least one wild-type *FUT1* allele can form the A or B antigen. As a result, individuals whose genotypes include the I^A or I^B allele and who have no wild-type *FUT1* allele are of the type O phenotype, regardless of their potential to make either antigen. An example of the outcome of matings between individuals heterozygous at both loci is illustrated in **Figure 4.6**. If many such



$I^AI^BHh \times I^AI^BHh$
Consideration of blood types
I^AI^B × I^AI^B
I^AI^A → 1/4 type A
I^AI^B
I^BI^A → 2/4 type AB
I^BI^B → 1/4 type B
Genotypes
Phenotypes
Consideration of H substance
Hh × Hh
HH
Hh
hH → 3/4 form H substance
hh → 1/4 do not form H substance
Genotypes
Phenotypes
Consideration of both gene pairs together
Of all offspring
Of all offspring
Final probabilities
1/4 type A
3/4 form H substance → 3/16 type A
1/4 do not form H substance → 1/16 type O
2/4 type AB
3/4 form H substance → 6/16 type AB
1/4 do not form H substance → 2/16 type O
1/4 type B
3/4 form H substance → 3/16 type B
1/4 do not form H substance → 1/16 type O
Final phenotypic ratio = 3/16 A: 6/16 AB: 3/16 B: 4/16 O

FIGURE 4.6 The outcome of a mating between individuals heterozygous at two genes determining their ABO blood type. Final phenotypes are calculated by considering each gene separately and then combining the results using the forked-line method.

individuals have children, the phenotypic ratio of 3A: 6AB: 3B: 4O is expected in their offspring.

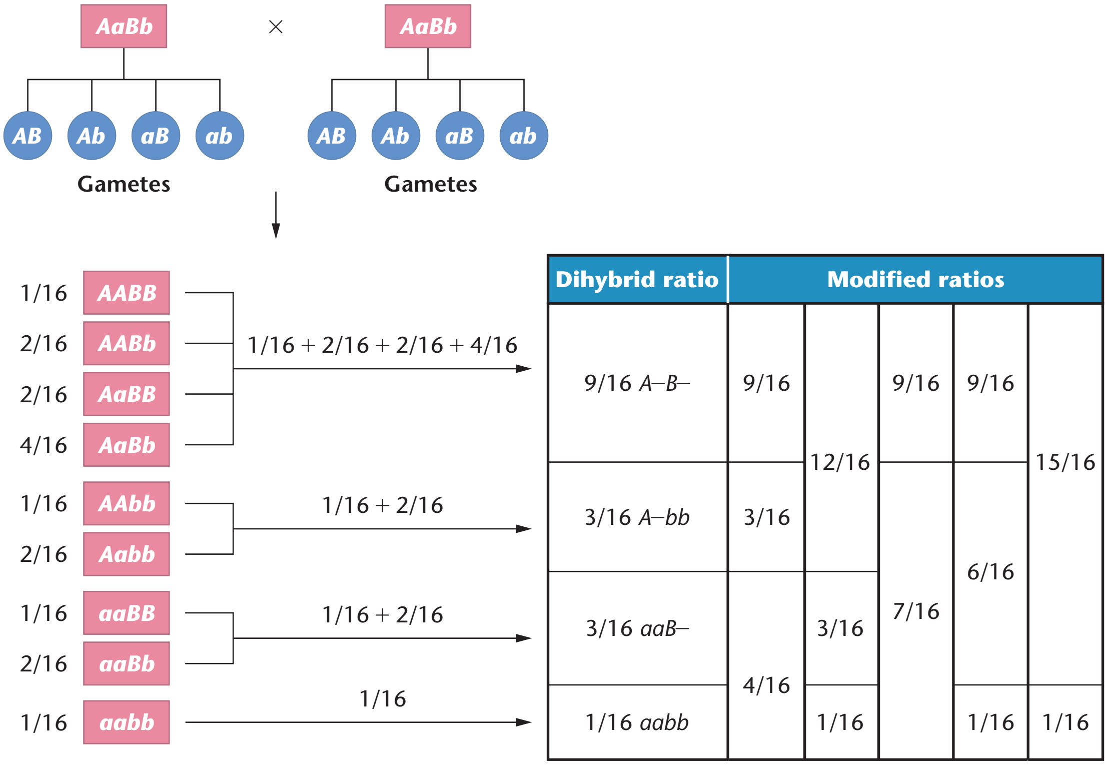
It is important to note two things when examining this cross and the predicted phenotypic ratio:

1. A key distinction exists between this cross and the modified dihybrid cross shown in Figure 4.5: *only one character—blood type—is being followed.* In the modified dihybrid cross in Figure 4.5, blood type *and* skin pigmentation are followed as separate phenotypic characteristics.
2. Even though only a single character was followed, the phenotypic ratio comes out in sixteenths. If we knew nothing about the H substance and the gene controlling it, we could still be confident (because the proportions are in sixteenths) that a second gene pair, other than that controlling the A and B antigens, was involved in the phenotypic expression. *When a single character is being studied, a ratio that is expressed in 16 parts (e.g., 3:6:3:4) suggests that two gene pairs are "interacting" in the expression of the phenotype under consideration.*

The study of gene interaction reveals a number of inheritance patterns that are modifications of the Mendelian dihybrid F_2 ratio (9:3:3:1). In several of the subsequent examples, epistasis has the effect of combining one or more of the four phenotypic categories in various ways. The generation of these four groups is reviewed in **Figure 4.7**, along with several modified ratios.

As we discuss these and other examples (see **Figure 4.8**), we will make several assumptions and adopt certain conventions:

1. In each case, distinct phenotypic classes are produced, each clearly discernible from all others. Such traits illustrate discontinuous variation, where phenotypic categories are discrete and qualitatively different from one another.
2. The genes considered in each cross are on different chromosomes and therefore assort independently of one another during gamete formation. To allow you to easily compare the results of different crosses, we designated alleles as *A, a* and *B, b* in each case.
3. When we assume that complete dominance exists within a gene pair, such that *AA* and *Aa* or *BB* and *Bb* are equivalent in their genetic effects, we use the designations *A*– or *B*– for both combinations, where the dash (–) indicates that either allele may be present without consequence to the phenotype.
4. All P_1 crosses involve homozygous individuals (e.g., *AABB* × *aabb*, *AAbb* × *aaBB* × *aaBB* × *AAbb*). Therefore, each F_1 generation consists of only heterozygotes of genotype *AaBb*.
5. In each example, the F_2 generation produced from these heterozygous parents is our main focus of analysis. When two genes are involved (Figure 4.7), the F_2 genotypes fall into four categories: 9/16 *A*–*B*–, 3/16 *A*–*bb*, 3/16 *aaB*–,



Dihybrid ratio	Modified ratios				
9/16 *A–B–*	9/16	12/16	9/16	9/16	15/16
3/16 *A–bb*	3/16		7/16	6/16	
3/16 *aaB–*	4/16	3/16			
1/16 *aabb*		1/16		1/16	1/16

FIGURE 4.7 Generation of various modified dihybrid ratios from the nine unique genotypes produced in a cross between individuals heterozygous at two genes.

Case	Organism	Character	F_2 Phenotypes				Modified ratio
			9/16	3/16	3/16	1/16	
1	Mouse	Coat color	agouti	albino	black	albino	9:3:4
2	Squash	Color	white		yellow	green	12:3:1
3	Pea	Flower color	purple	white			9:7
4	Squash	Fruit shape	disc	sphere		long	9:6:1
5	Chicken	Color	white		colored	white	13:3
6	Mouse	Color	white-spotted	white	colored	white-spotted	10:3:3
7	Shepherd's purse	Seed capsule	triangular			ovoid	15:1
8	Flour beetle	Color	6/16 sooty and 3/16 red	black	jet	black	6:3:3:4

FIGURE 4.8 The basis of modified dihybrid F_2 phenotypic ratios resulting from crosses between doubly heterozygous F_1 individuals. The four groupings of the F_2 genotypes shown in Figure 4.7 and across the top of this figure are combined in various ways to produce these ratios.

and 1/16 *aabb*. Because of dominance, all genotypes in each category are equivalent in their effect on the phenotype.

Case 1 is the inheritance of coat color in mice (Figure 4.8). Normal wild-type coat color is agouti, a grayish pattern formed by alternating bands of pigment on each hair (see Figure 4.4). Agouti is dominant to black (nonagouti) hair, which results from the homozygous expression of a recessive mutation that we designate *a*. Thus, *A*– results in agouti, whereas *aa* yields black coat color. When a recessive mutation, *b*, at a separate locus is homozygous, it eliminates pigmentation altogether, yielding albino mice (*bb*), regardless of the genotype at the *a* locus. Thus, in a cross between agouti (*AABB*) and albino (*aabb*) parents, members of the F_1 are all *AaBb* and have agouti coat color. In the F_2 progeny of a cross between two F_1 double heterozygotes, the following genotypes and phenotypes are observed:

$$F_1\text{: } AaBb \times AaBb$$
$$\downarrow$$

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A*–*B*–	agouti	9/16 agouti 3/16 black 4/16 albino
3/16	*A*–*bb*	albino	
3/16	*aaB*–	black	
1/16	*aabb*	albino	

We can envision gene interaction yielding the observed 9:3:4 F_2 ratio as a two-step process:

$$\text{Precursor molecule (colorless)} \xrightarrow[B-]{\text{Gene B} \downarrow} \text{Black pigment} \xrightarrow[A-]{\text{Gene A} \downarrow} \text{Agouti pattern}$$

In the presence of a *B* allele, black pigment can be made from a colorless substance. In the presence of an *A* allele, the black pigment is deposited during the development of hair in a pattern that produces the agouti phenotype. If the *aa* genotype occurs, all of the hair remains black. If the *bb* genotype occurs, no black pigment is produced, regardless of the presence of the *A* or *a* alleles, and the mouse is albino. Therefore, the *bb* genotype masks or suppresses the expression of the *A* allele. As a result, this is referred to as *recessive epistasis.*

A second type of epistasis, called *dominant epistasis,* occurs when a dominant allele at one genetic locus masks the expression of the alleles of a second locus. For instance, case 2 of Figure 4.8 deals with the inheritance of fruit color in summer squash. Here, the dominant allele *A* results in white fruit color regardless of the genotype at a second locus, *B*. In the absence of a dominant *A* allele (the *aa* genotype), *BB* or *Bb* results in yellow color, while *bb* results in green color. Therefore, if two white-colored double heterozygotes (*AaBb*) are crossed, this type of epistasis generates an interesting phenotypic ratio:

$$F_1\text{: } AaBb \quad \times \quad AaBb$$
$$\downarrow$$

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A–B–*	white	12/16 white 3/16 yellow 1/16 green
3/16	*A–bb*	white	
3/16	*aaB–*	yellow	
1/16	*aabb*	green	

Of the offspring, 9/16 are *A–B–* and are thus white. The 3/16 bearing the genotypes *A–bb* are also white. Of the remaining squash, 3/16 are yellow (*aaB–*), while 1/16 are green (*aabb*). Thus, the modified phenotypic ratio of 12:3:1 occurs.

Our third example (case 3 of Figure 4.8), first discovered by William Bateson and Reginald Punnett (of Punnett square fame), is demonstrated in a cross between two true-breeding strains of white-flowered sweet peas. Unexpectedly, the results of this cross yield all purple F_1 plants, and the F_2 plants occur in a ratio of 9/16 purple to 7/16 white. The proposed explanation suggests that the presence of at least one dominant allele of each of two genes is essential in order for flowers to be purple. Thus, this cross represents a case of *complementary gene interaction.* All other genotype combinations yield white flowers because the homozygous condition of either recessive allele masks the expression of the dominant allele at the other locus.

The cross is shown as follows:

$$\begin{array}{ccc} P_1\text{: } AAbb & \times & aaBB \\ \text{white} & & \text{white} \\ & \downarrow & \\ F_1\text{: All } AaBb & & \text{(purple)} \end{array}$$

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A–B–*	purple	9/16 purple 7/16 white
3/16	*A–bb*	white	
3/16	*aaB–*	white	
1/16	*aabb*	white	

We can now envision how two genes might yield such results:

	Gene A		Gene B	
Precursor substance (colorless)	$\downarrow$ $\longrightarrow$ *A–*	Intermediate product (colorless)	$\downarrow$ $\longrightarrow$ *B–*	Final product (purple)

At least one dominant allele from each gene is necessary to ensure both biochemical conversions to the final product, yielding purple flowers. In the preceding cross, this will occur in 9/16 of the F_2 offspring. All other plants (7/16) have flowers that remain white.

These three examples illustrate in a simple way how the products of two genes interact to influence the development of a common phenotype. In other instances, more than two genes and their products are involved in controlling phenotypic expression.

Novel Phenotypes

Other cases of gene interaction yield novel, or new, phenotypes in the F_2 generation, in addition to producing modified dihybrid ratios. Case 4 in Figure 4.8 depicts the inheritance of fruit shape in the summer squash *Cucurbita pepo.* When plants with disc-shaped fruit (*AABB*) are crossed with plants with long fruit (*aabb*), the F_1 generation all have disc fruit. However, in the F_2 progeny, fruit with a novel shape—sphere—appear, as well as fruit exhibiting the parental phenotypes. A variety of fruit shapes are shown in **Figure 4.9**.

The F_2 generation, with a modified 9:6:1 ratio, is generated as follows:

$$\begin{array}{ccc} F_1\text{: } AaBb & \times & AaBb \\ \text{disc} & & \text{disc} \\ & \downarrow & \end{array}$$

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A–B–*	disc	9/16 disc 6/16 sphere 1/16 long
3/16	*A–bb*	sphere	
3/16	*aaB–*	sphere	
1/16	*aabb*	long	

In this example of gene interaction, both gene pairs influence fruit shape equally. A dominant allele at either locus ensures a sphere-shaped fruit. In the absence of dominant alleles, the fruit is long. However, if both dominant alleles (*A* and *B*) are present, the fruit displays a flattened, disc shape.

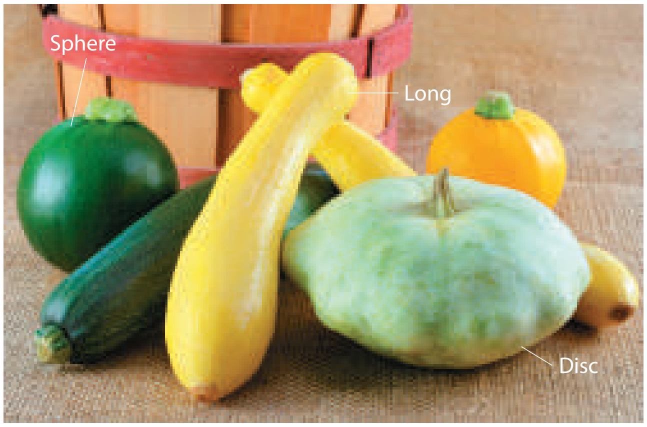


FIGURE 4.9 Summer squash exhibiting various fruit-shape phenotypes, including disc, long, and sphere.

NOW SOLVE THIS

4.2 In some plants a red flower pigment, cyanidin, is synthesized from a colorless precursor. The addition of a hydroxyl group (OH^-) to the cyanidin molecule causes it to become purple. In a cross between two randomly selected purple varieties, the following results were obtained:

94 purple
31 red
43 white

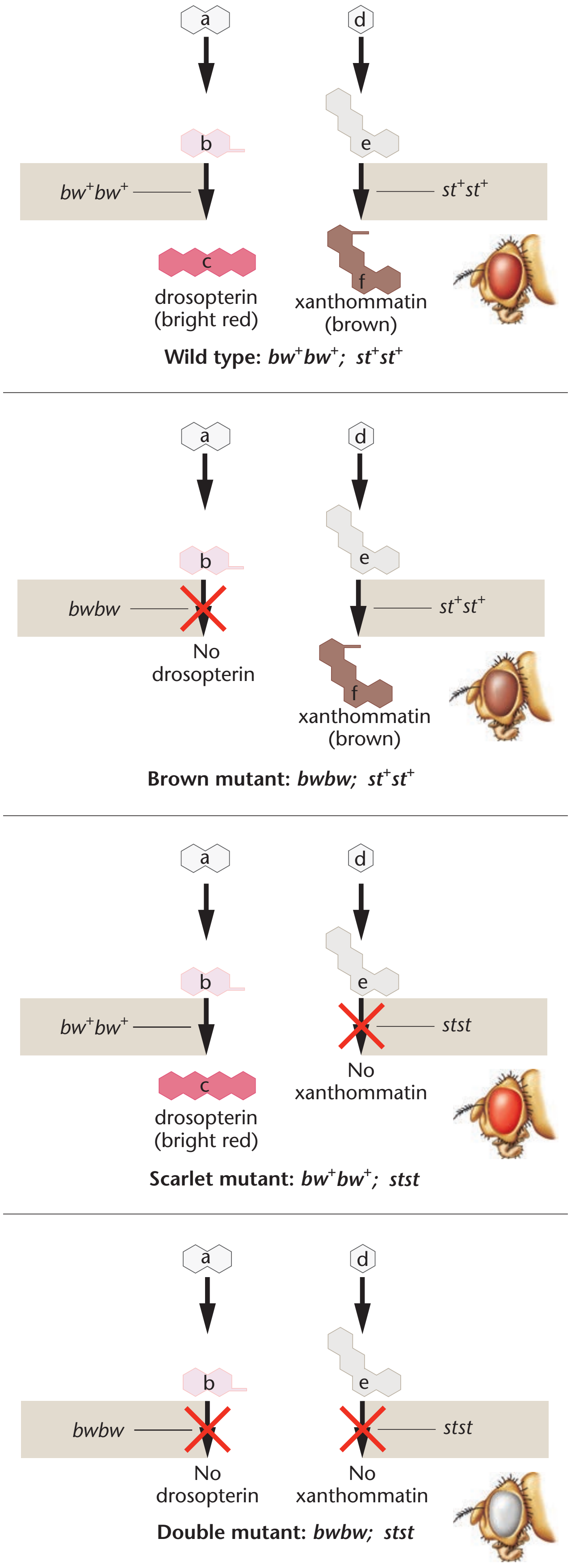
How many genes are involved in the determination of these flower colors? Which genotypic combinations produce which phenotypes? Diagram the purple × purple cross.

- **HINT:** *This problem describes a plant in which flower color, a single characteristic, can take on one of three variations. The key to its solution is to first analyze the raw data and convert the numbers to a meaningful ratio. This will guide you in determining how many gene pairs are involved. Then you can group the genotypes in a way that corresponds to the phenotypic ratio.*

For more practice, see Problems 15–18.

Another interesting example of an unexpected phenotype arising in the F_2 generation is the inheritance of eye color in *Drosophila melanogaster*. As mentioned earlier, the wild-type eye color is brick red. When two autosomal recessive mutants, *brown* and *scarlet*, are crossed, the F_1 generation consists of flies with wild-type eye color. In the F_2 generation, wild, scarlet, brown, and white-eyed flies are found in a 9:3:3:1 ratio. While this ratio is numerically the same as Mendel's dihybrid ratio, the *Drosophila* cross involves only one character: eye color. This is an important distinction to make when modified dihybrid ratios resulting from gene interaction are studied.

The *Drosophila* cross is an excellent example of gene interaction because the biochemical basis of eye color in this organism has been determined (**Figure 4.10**). *Drosophila*, as a typical arthropod, has compound eyes made up of hundreds of individual visual units called ommatidia. The wild-type eye color is due to the deposition and mixing



Wild type: bw^+bw^+; st^+st^+

Brown mutant: $bwbw$; st^+st^+

Scarlet mutant: bw^+bw^+; $stst$

Double mutant: $bwbw$; $stst$

FIGURE 4.10 A theoretical explanation of the biochemical basis of the four eye color phenotypes produced in a cross between *Drosophila* with brown eyes and scarlet eyes. In the presence of at least one wild-type bw^+ allele, an enzyme is produced that converts substance b to c, and the pigment drosopterin is synthesized. In the presence of at least one wild-type st^+ allele, substance e is converted to f, and the pigment xanthommatin is synthesized. The homozygous presence of the recessive *st* or *bw* mutant allele blocks the synthesis of the respective pigment molecule. Either one, both, or neither of these pathways can be blocked, depending on the genotype.

of two separate pigment groups in each ommatidium—the bright-red **drosopterins** and the brown **xanthommatins.** Each type of pigment is produced by a separate biosynthetic pathway. Each step of each pathway is catalyzed by a separate enzyme and is thus under the control of a separate gene. As shown in Figure 4.10, the *brown* mutation, when homozygous, interrupts the pathway leading to the synthesis of the bright-red pigments. Because only xanthommatin pigments are present, the eye is brown. The *scarlet* mutation, affecting a gene located on a separate autosome, interrupts the pathway leading to the synthesis of the brown xanthommatins and renders the eye color bright red in homozygous mutant flies. Each mutation apparently causes the production of a nonfunctional enzyme. Flies that are double mutants and thus homozygous for both *brown* and *scarlet* lack both functional enzymes and can make neither of the pigments; they represent the novel white-eyed flies appearing in 1/16 of the F_2 generation. Note that the absence of pigment in these flies is not due to the X-linked *white* mutation, in which pigments can be synthesized but the necessary precursors cannot be transported into the cells making up the ommatidia.

Other Modified Dihybrid Ratios

The remaining cases (5–8) in Figure 4.8 illustrate additional modifications of the dihybrid ratio and provide still other examples of gene interactions. As you will note, ratios of 13:3, 10:3:3; 15:1, and 6:3:3:4 are illustrated. These cases, like the four preceding them, have two things in common. First, we need not violate the principles of segregation and independent assortment to explain the inheritance pattern of each case. Therefore, the added complexity of inheritance in these examples does not detract from the validity of Mendel's conclusions. Second, the F_2 phenotypic ratio in each example has been expressed in sixteenths. When sixteenths are seen in the ratios of crosses where the inheritance pattern is unknown, they suggest to geneticists that two gene pairs are controlling the observed phenotypes. You should make the same inference in your analysis of genetics problems. Other insights into solving genetics problems are provided in the *Insights and Solutions* section at the conclusion of this chapter.

4.9 Complementation Analysis Can Determine if Two Mutations Causing a Similar Phenotype Are Alleles of the Same Gene

An interesting situation arises when two mutations that both produce a similar phenotype are isolated independently. Suppose that two investigators independently isolate and establish a true-breeding strain of wingless *Drosophila* and demonstrate that each mutant phenotype is due to a recessive mutation. We might assume that both strains contain mutations in the same gene. However, since we know that many genes are involved in the formation of wings, we must consider the possibility that mutations in any one of them might inhibit wing formation during development. This is the case with any *heterogeneous trait,* a concept introduced earlier in this chapter in our discussion of hereditary deafness. An analytical procedure called **complementation analysis** allows us to determine whether two independently isolated mutations are in the same gene—that is, whether they are alleles—or whether they represent mutations in separate genes.

To repeat, our analysis seeks to answer this simple question: *Are two mutations that yield similar phenotypes present in the same gene or in two different genes?* To find the answer, we cross the two mutant strains and analyze the F_1 generation. The two possible alternative outcomes and their interpretations are shown in Figure 4.11. To discuss these possibilities (case 1 and case 2), we designate one of the mutations m^a and the other m^b.

Case 1. *All offspring develop normal wings.*
Interpretation: The two recessive mutations are in separate genes and are not alleles of one another. Following the cross, all F_1 flies are heterozygous for both genes. Since each mutation is in a separate gene and each F_1 fly is heterozygous at both loci, the normal products of both genes are produced (by the one normal copy of each gene), and wings develop. Under such circumstances, the genes complement one another in restoration of the wild-type phenotype, and complementation is said to occur because the two mutations are in different genes.

Case 2. *All offspring fail to develop wings.*
Interpretation: The two mutations affect the same gene and are alleles of one another. Complementation does not occur. Since the two mutations affect the same gene, the F_1 flies are homozygous for the two mutant alleles (the m^a allele and the m^b allele). No normal product of the gene is produced, and in the absence of this essential product, wings do not form.

Complementation analysis, as originally devised by the *Drosophila* geneticist Edward B. Lewis, may be used to screen any number of individual mutations that result in the same phenotype. Such an analysis may reveal that only a single gene is involved or that two or more genes are involved. All mutations determined to be present in any single gene are said to fall into the same **complementation group,** and they will complement mutations in all other groups. When large numbers of

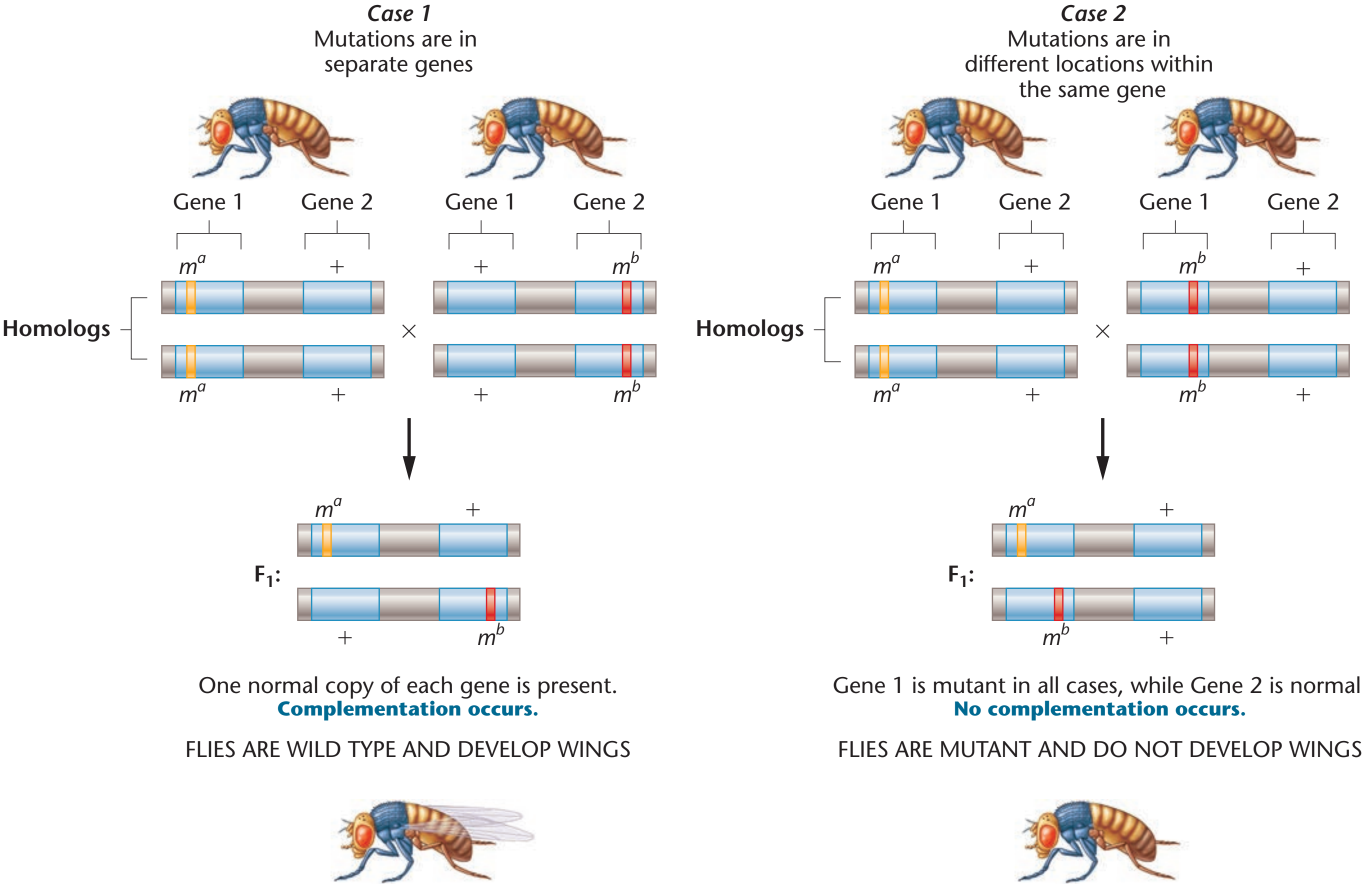

Case 1
Mutations are in separate genes
Case 2
Mutations are in different locations within the same gene
Gene 1
Gene 2
m^a
m^b
+
Homologs
×
F_1:
One normal copy of each gene is present.
Complementation occurs.
FLIES ARE WILD TYPE AND DEVELOP WINGS
Gene 1 is mutant in all cases, while Gene 2 is normal
No complementation occurs.
FLIES ARE MUTANT AND DO NOT DEVELOP WINGS

FIGURE 4.11 Complementation analysis of alternative outcomes of two wingless mutations in *Drosophila* (m^a and m^b). In case 1, the mutations are not alleles of the same gene, while in case 2, the mutations are alleles of the same gene.

mutations affecting the same trait are available and studied using complementation analysis, it is possible to predict the total number of genes involved in the determination of that trait.

4.10 Expression of a Single Gene May Have Multiple Effects

While the previous sections have focused on the effects of two or more genes on a single characteristic, the converse situation, where expression of a single gene has multiple phenotypic effects, is also quite common. This phenomenon, which often becomes apparent when phenotypes are examined carefully, is referred to as **pleiotropy.** Many excellent examples can be drawn from human disorders, and we will review two such cases to illustrate this point.

The first disorder is **Marfan syndrome,** a human malady resulting from an autosomal dominant mutation in the gene encoding the connective tissue protein *fibrillin.* Because this protein is widespread in many tissues in the body, one would expect multiple effects of such a defect. In fact, fibrillin is important to the structural integrity of the lens of the eye, to the lining of vessels such as the aorta, and to bones, among other tissues. As a result, the phenotype associated with Marfan syndrome includes lens dislocation, increased risk of aortic aneurysm, and lengthened long bones in limbs. This disorder is of historical interest in that speculation abounds that Abraham Lincoln was afflicted.

A second example involves another human autosomal dominant disorder, **porphyria variegata.** Afflicted individuals cannot adequately metabolize the porphyrin component of hemoglobin when this respiratory pigment is broken down as red blood cells are replaced. The accumulation of excess porphyrins is immediately evident in the urine, which takes on a deep red color. However, this phenotypic characteristic is merely diagnostic. The severe features of the disorder are due to the toxicity of the buildup of porphyrins in the body, particularly in the brain. Complete phenotypic characterization includes abdominal pain, muscular weakness, fever, a racing pulse, insomnia, headaches, vision problems (that can lead to blindness), delirium, and ultimately convulsions. As you can see, deciding which phenotypic trait best characterizes the disorder is impossible.

Like Marfan syndrome, porphyria variegata is also of historical significance. George III, King of England during the American Revolution, is believed to have suffered from episodes involving all of the above symptoms. He ultimately became blind and senile prior to his death.

We could cite many other examples to illustrate pleiotropy, but suffice it to say that if one looks carefully, most mutations display more than a single manifestation when expressed.

4.11 X-Linkage Describes Genes on the X Chromosome

In many animals and some plant species, one of the sexes contains a pair of unlike chromosomes that are involved in sex determination. In many cases, these are designated as X and Y. For example, in both *Drosophila* and humans, males contain an X and a Y chromosome, whereas females contain two X chromosomes. The Y chromosome must contain a region of pairing homology with the X chromosome if the two are to synapse and segregate during meiosis, but a major portion of the Y chromosome in humans as well as other species is considered to be relatively inert genetically. While we now recognize a number of male-specific genes on the human Y chromosome, it lacks copies of most genes present on the X chromosome. As a result, genes present on the X chromosome exhibit patterns of inheritance that are very different from those seen with autosomal genes. The term **X-linkage** is used to describe these situations.

In the following discussion, we will focus on inheritance patterns resulting from genes present on the X but absent from the Y chromosome. This situation results in a modification of Mendelian ratios, the central theme of this chapter.

X-Linkage in *Drosophila*

One of the first cases of X-linkage was documented in 1910 by Thomas H. Morgan during his studies of the *white* eye mutation in *Drosophila* (**Figure 4.12**). The normal wild-type red eye color is dominant to white eye color.

Morgan's work established that the inheritance pattern of the white-eye trait was clearly related to the sex of the parent carrying the mutant allele. Unlike the outcome of the typical Mendelian monohybrid cross where F_1 and F_2 data were similar regardless of which P_1 parent exhibited the recessive mutant trait, reciprocal crosses between white-eyed and red-eyed flies did not yield identical results. Morgan's analysis led to the conclusion that the *white* locus is present on the X chromosome rather than on one of the autosomes. Both the gene and the trait are said to be X-linked.

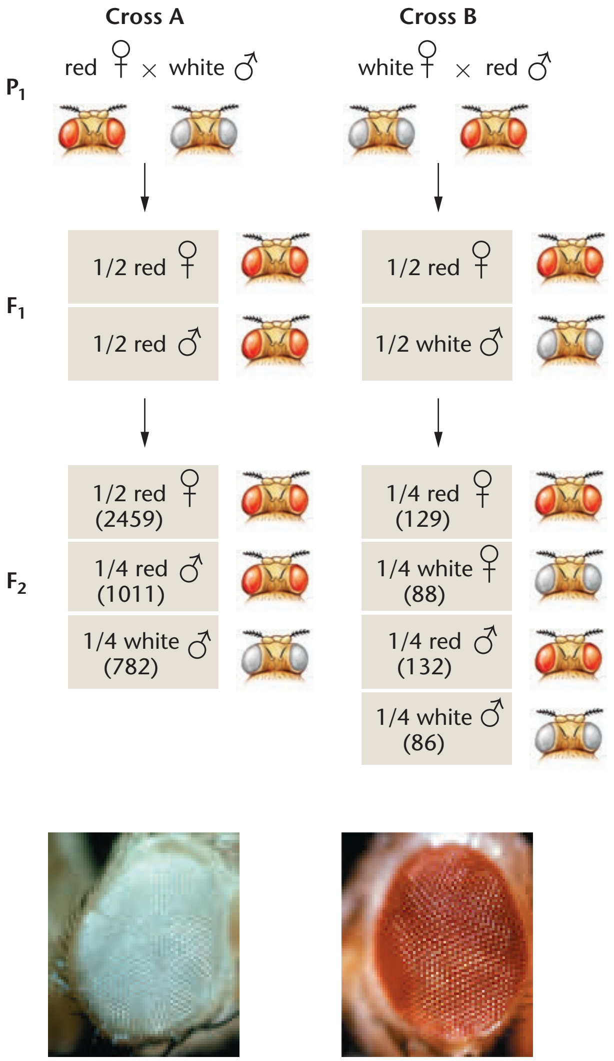


FIGURE 4.12 The F_1 and F_2 results of T. H. Morgan's reciprocal crosses involving the X-linked *white* mutation in *Drosophila melanogaster*. The actual data are shown in parentheses. The photographs show white eye and the brick-red wild-type eye color.

Results of reciprocal crosses between white-eyed and red-eyed flies are shown in Figure 4.12. The obvious differences in phenotypic ratios in both the F_1 and F_2 generations are dependent on whether or not the P_1 white-eyed parent was male or female.

Morgan was able to correlate these observations with the difference found in the sex-chromosome composition of male and female *Drosophila*. He hypothesized that the recessive allele for white eye is found on the X chromosome, but its corresponding locus is absent from the Y chromosome. Females thus have two available gene loci, one on each X chromosome, whereas males have only one available locus, on their single X chromosome.

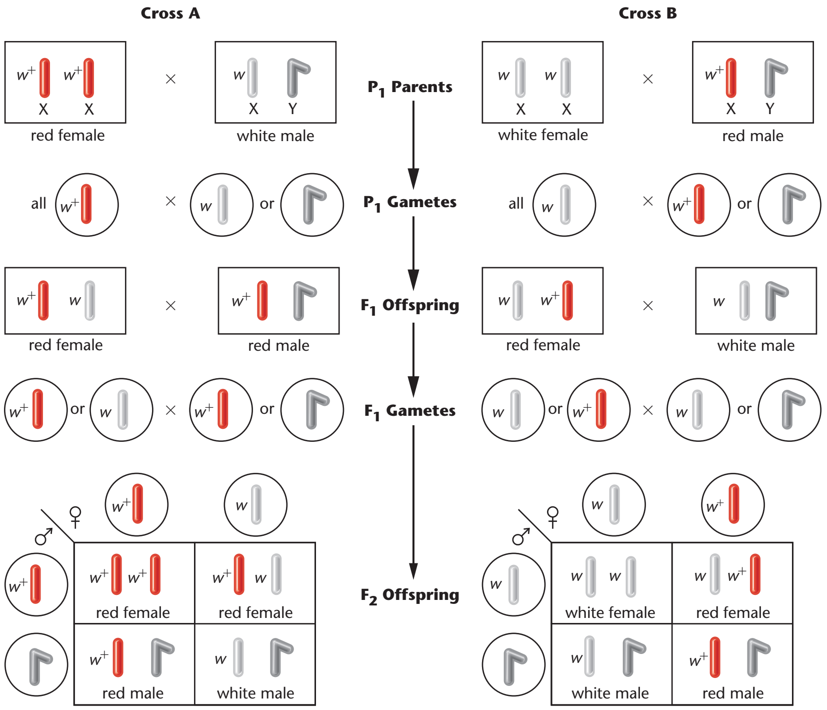


FIGURE 4.13 The chromosomal explanation of the results of the X-linked crosses shown in Figure 4.12.

Morgan's interpretation of X-linked inheritance, shown in Figure 4.13, provides a suitable theoretical explanation for his results. Since the Y chromosome lacks homology with almost all genes on the X chromosome, these alleles present on the X chromosome of the males will be directly expressed in the phenotype. Males cannot be either homozygous or heterozygous for X-linked genes; instead, their condition— possession of only one copy of a gene in an otherwise diploid cell—is referred to as **hemizygosity.** The individual is said to be **hemizygous.** One result of X-linkage is the *crisscross pattern of inheritance,* in which phenotypic traits controlled by recessive X-linked genes are passed from homozygous mothers to all sons. This pattern occurs because females exhibiting a recessive trait must contain the mutant allele on both X chromosomes. Because male offspring receive one of their mother's two X chromosomes and are hemizygous for all alleles present on that X, all sons will express the same recessive X-linked traits as their mother.

Morgan's work has taken on great historical significance. By 1910, the correlation between Mendel's work and the behavior of chromosomes during meiosis had provided the basis for the **chromosome theory of inheritance.** Morgan's work, and subsequently that of his student Calvin Bridges around 1920, provided direct evidence that genes are transmitted on specific chromosomes and is considered the first solid experimental evidence in support of this theory.

X-Linkage in Humans

In humans, many genes and the respective traits controlled by them are recognized as being linked to the X chromosome (see **Table 4.2**). These X-linked traits can be easily identified in a pedigree because of the crisscross pattern of inheritance. A pedigree for one form of human **color blindness** is shown in Figure 4.14. The mother in generation I passes the trait to all her sons but to none of her daughters. If the offspring in generation II marry normal individuals, the color-blind sons will produce all normal male and female offspring (III-1, 2, and 3); the normal-visioned daughters will produce normal-visioned female offspring (III-4, 6, and 7), as well as color-blind (III-8) and normal-visioned (III-5) male offspring.

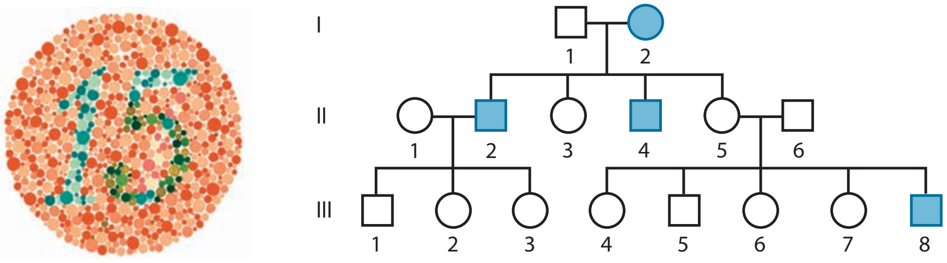


FIGURE 4.14 A human pedigree of the X-linked color-blindness trait. The photograph is of an Ishihara color-blindness chart, which tests for red–green color blindness. Those with normal vision will see the number 15, while those with red–green color blindness will see the number 17.

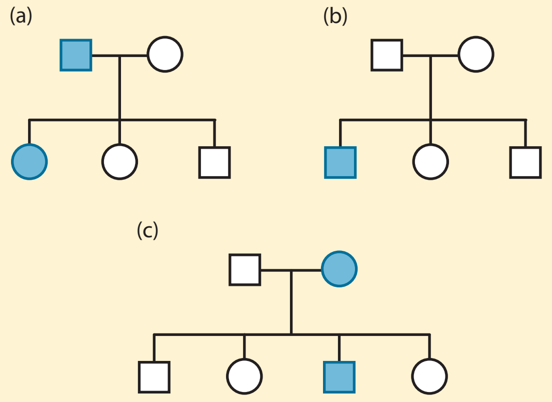
TABLE 4.2 Human X-Linked Traits

Condition	Characteristics
Color blindness, deutan type	Insensitivity to green light
G-6-PD deficiency	Deficiency of glucose-6-phosphate dehydrogenase; severe anemic reaction following intake of primaquines in drugs and certain foods, including fava beans
Hemophilia A	Classic form of clotting deficiency; deficiency of clotting factor VIII
Hemophilia B	Christmas disease; deficiency of clotting factor IX
Lesch–Nyhan syndrome	Deficiency of hypoxanthine-guanine phosphoribosyl transferase enzyme (HPRT) leading to motor and mental retardation, self-mutilation, and early death
Duchenne muscular dystrophy	Progressive, life-shortening disorder characterized by muscle degeneration and weakness; deficiency of the protein dystrophin

The way in which X-linked genes are transmitted causes unusual circumstances associated with recessive X-linked disorders, in comparison to recessive autosomal disorders. For example, if an X-linked disorder debilitates or is lethal to the affected individual prior to reproductive maturation, the disorder occurs exclusively in males. This is so because the only sources of the lethal allele in the population are in heterozygous females who are "carriers" and do not express the disorder. They pass the allele to one-half of their sons, who develop the disorder because they are hemizygous but rarely, if ever, reproduce. Heterozygous females also pass the allele to one-half of their daughters, who become carriers but do not develop the disorder. An example of such an X-linked disorder is *Duchenne muscular dystrophy*. The disease has an onset prior to age 6 and is often lethal around age 20. It normally occurs only in males.

NOW SOLVE THIS

4.3 Below are three pedigrees. For each trait, consider whether it is or is not consistent with X-linked recessive inheritance. In a sentence or two, indicate why or why not.



■ **HINT:** *This problem involves potential X-linked recessive traits as analyzed in pedigrees. The key to its solution is to focus on hemizygosity, where an X-linked recessive allele is always expressed in males, but never passed from a father to his sons. Homozygous females, on the other hand, pass the trait to all sons, but not to their daughters unless the father is also affected.*

For more practice, see Problems 30 and 40.

4.12 In Sex-Limited and Sex-Influenced Inheritance, an Individual's Sex Influences the Phenotype

In contrast to X-linked inheritance, patterns of gene expression may be affected by the sex of an individual even when the genes are not on the X chromosome. In numerous examples in different organisms, the sex of the individual plays a determining role in the expression of a phenotype. In some cases, the expression of a specific phenotype is absolutely limited to one sex; in others, the sex of an individual influences the expression of a phenotype that is not limited to one sex or the other. This distinction differentiates **sex-limited inheritance** from **sex-influenced inheritance.**

In both types of inheritance, autosomal genes are responsible for the existence of contrasting phenotypes, but the expression of these genes is dependent on the hormone constitution of the individual. Thus, the heterozygous genotype may exhibit one phenotype in males and the contrasting one in females. In domestic fowl, for example, tail and neck plumage is often distinctly different in males and females (**Figure 4.15**), demonstrating *sex-limited inheritance*. Cock feathering is longer, more curved, and pointed, whereas hen feathering is shorter and less curved. Inheritance of these feather phenotypes is controlled by a single pair of autosomal alleles whose expression

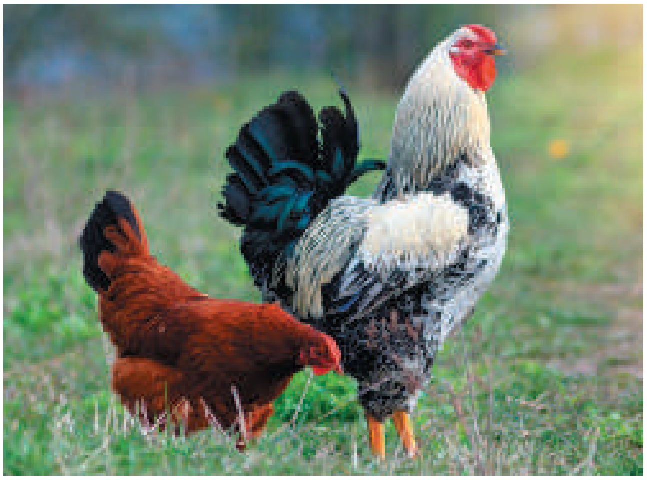


FIGURE 4.15 Hen feathering (left) and cock feathering (right) in domestic fowl. The hen's feathers are shorter and less curved.

is modified by the individual's sex hormones. As shown in the following chart, hen feathering is due to a dominant allele, *H,* but regardless of the homozygous presence of the recessive *h* allele, all females remain hen-feathered. Only in males does the *hh* genotype result in cock feathering.

Genotype	Phenotype	
	♀	♂
HH	Hen-feathered	Hen-feathered
Hh	Hen-feathered	Hen-feathered
hh	Hen-feathered	Cock-feathered

In certain breeds of fowl, the hen feathering or cock feathering allele has become fixed in the population. In the Leghorn breed, all individuals are of the *hh* genotype; as a result, males always differ from females in their plumage. Seabright bantams are all *HH,* showing no sexual distinction in feathering phenotypes.

Another example of sex-limited inheritance involves the autosomal genes responsible for milk yield in dairy cattle. Regardless of the overall genotype that influences the quantity of milk production, those genes are obviously expressed only in females.

Cases of *sex-influenced inheritance* include pattern baldness in humans, horn formation in certain breeds of sheep (e.g., Dorset Horn sheep), and certain coat patterns in cattle. In such cases, autosomal genes are responsible for the contrasting phenotypes, and while the trait may be displayed by both males and females, the expression of these genes is dependent on the hormone constitution of the individual. Thus, the heterozygous genotype exhibits one phenotype in one sex and the contrasting one in the other. For example, pattern baldness in humans, where the hair is very thin or absent on the top of the head (**Figure 4.16**), is inherited in the following way:

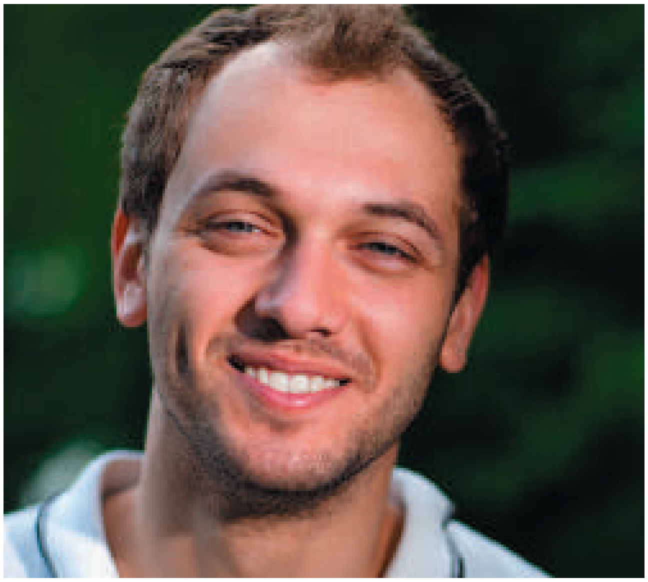


FIGURE 4.16 Pattern baldness, a sex-influenced autosomal trait, in a young man.

Genotype	Phenotype	
	♀	♂
BB	Bald	Bald
Bb	Not bald	Bald
bb	Not bald	Not bald

Females can display pattern baldness, but this phenotype is much more prevalent in males. When females do inherit the *BB* genotype, the phenotype is less pronounced than in males and is expressed later in life.

4.13 Genetic Background and the Environment May Alter Phenotypic Expression

In the final section of this chapter we consider *phenotypic expression.* We assumed that the genotype of an organism is always directly expressed in its phenotype (Chapters 2 and 3). For example, pea plants homozygous for the recessive *d* allele (*dd*) will always be dwarf. We discussed gene expression as though the genes operate in a closed system in which the presence or absence of functional products directly determines the collective phenotype of an individual. The situation is actually much more complex. Most gene products function within the internal milieu of the cell, and cells interact with one another in various ways. Furthermore, the organism exists under diverse environmental influences. Thus, gene expression and the resultant phenotype are often modified through the interaction between an individual's particular genotype and the external environment. In this final section of this chapter, we will deal with some of the variables that are known to modify gene expression.

Penetrance and Expressivity

Some mutant genotypes are always expressed as a distinct phenotype, whereas others produce a proportion of individuals whose phenotypes cannot be distinguished from normal (wild type). The degree of expression of a particular trait can be studied quantitatively by determining the *penetrance* and *expressivity* of the genotype under investigation.

The percentage of individuals that show at least some degree of expression of a mutant genotype defines the **penetrance** of the mutation. For example, the phenotypic expression of many of the mutant alleles found in *Drosophila* can overlap with wild-type expression. If 15 percent of flies with a given mutant genotype show the wild-type appearance, the mutant gene is said to have a penetrance of 85 percent.

By contrast, **expressivity** reflects the *range of expression* of the mutant genotype. Flies homozygous for the recessive mutant gene *eyeless* exhibit phenotypes that range from the presence of normal eyes to a partial reduction in size to the complete absence of one or both eyes

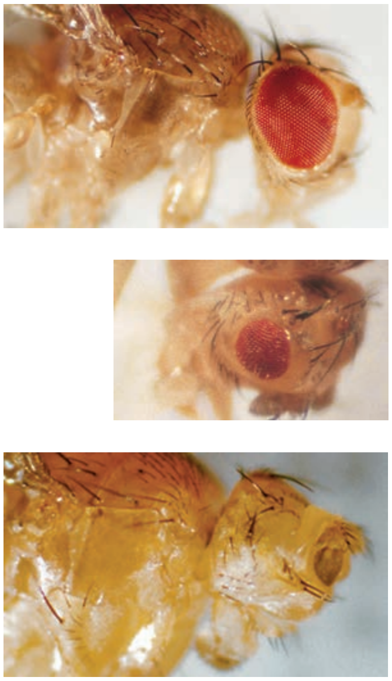


FIGURE 4.17 Variable expressivity as shown in flies homozygous for the *eyeless* mutation in *Drosophila*. Gradations in phenotype range from wild type to partial reduction to eyeless.

(Figure 4.17). Although the average reduction of eye size is one-fourth to one-half, expressivity ranges from complete loss of both eyes to completely normal eyes.

Examples such as the expression of the *eyeless* gene have provided the basis for experiments to determine the causes of phenotypic variation. If the laboratory environment is held constant and extensive variation is still observed, other genes may be influencing or modifying the phenotype. On the other hand, if the genetic background is not the cause of the phenotypic variation, environmental factors such as temperature, humidity, and nutrition may be involved. In the case of the eyeless phenotype, experiments have shown that both genetic background and environmental factors influence its expression.

Genetic Background: Position Effects

Although it is difficult to assess the specific effect of the **genetic background** and the expression of a gene responsible for determining a potential phenotype, one effect of genetic background has been well characterized, called the **position effect.** In such instances, the physical location of a gene in relation to other genetic material may influence its expression. For example, if a region of a chromosome is relocated or rearranged (called a translocation or inversion event), normal expression of genes in that chromosomal region may be modified. This is particularly true if the gene is relocated to or near certain areas of the chromosome that are condensed and genetically inert, referred to as **heterochromatin.**

An example of a position effect involves female *Drosophila* heterozygous for the X-linked recessive eye color mutant *white* (*w*). The w^+/w genotype normally results in a wild-type brick-red eye color. However, if the region of the X chromosome containing the wild-type w^+ allele is translocated so that it is close to a heterochromatic region, expression of the w^+ allele is modified. Instead of having a red color, the eyes are variegated, or mottled with red and white patches (Figure 4.18). Therefore, following translocation, the dominant effect of the normal w^+ allele is intermittent. A similar position effect is produced if a heterochromatic region is relocated next to the *white* locus on the X chromosome. Apparently, heterochromatic regions inhibit the expression of adjacent genes. Loci in many other organisms also exhibit position effects, providing proof that alteration of the normal arrangement of genetic information can modify its expression.

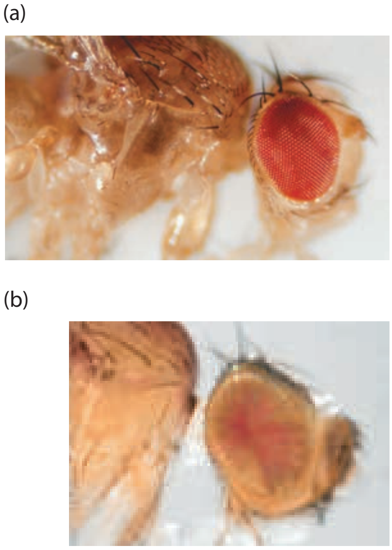


FIGURE 4.18 Position effect, as illustrated in the eye phenotype in two female *Drosophila* heterozygous for the *white* gene. (a) Normal dominant phenotype showing brick-red eye color. (b) Variegated color of an eye caused by translocation of the *white* gene to another location in the genome.

Temperature Effects—An Introduction to Conditional Mutations

Chemical activity depends on the kinetic energy of the reacting substances, which in turn depends on the surrounding temperature. We can thus expect temperature to influence phenotypes. An example is seen in the evening primrose, which produces red flowers when grown at 23°C and white flowers when grown at 18°C. An even more striking example is seen in Siamese cats and Himalayan rabbits, which exhibit dark fur in certain regions where their body temperature is slightly cooler, particularly the nose, ears, and paws (Figure 4.19). In these cases, it appears that the enzyme normally responsible for pigment production is functional only at the lower temperatures present in the extremities, but it loses its catalytic function at the slightly higher temperatures found throughout the rest of the body.

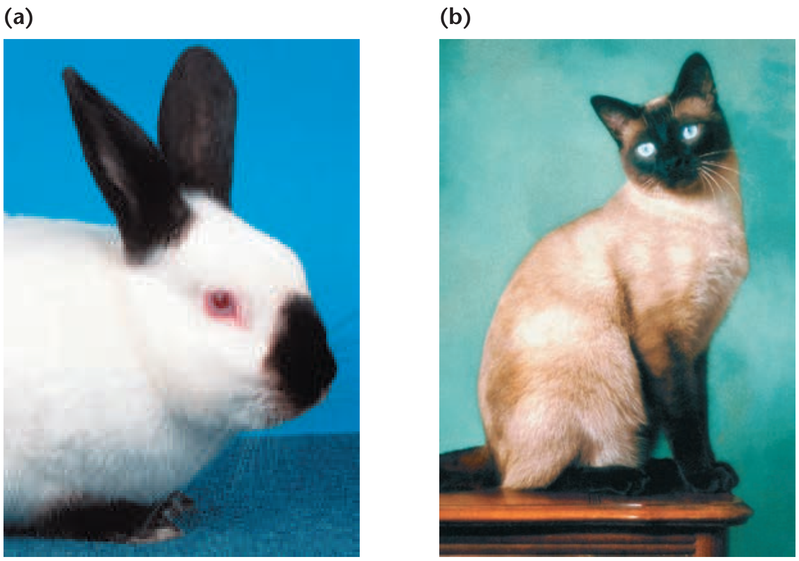


FIGURE 4.19 (a) A Himalayan rabbit. (b) A Siamese cat. Both show dark fur color on the snout, ears, and paws. These patches are due to the effect of a temperature-sensitive allele responsible for pigment production.

Mutations whose expression is affected by temperature, called **temperature-sensitive mutations,** are examples of **conditional mutations,** whereby phenotypic expression is determined by environmental conditions. Examples of temperature-sensitive mutations are known in viruses and a variety of organisms, including bacteria, fungi, and *Drosophila.* In extreme cases, an organism carrying a mutant allele may express a mutant phenotype when grown at one temperature but express the wild-type phenotype when reared at another temperature. This type of temperature effect is useful in studying mutations that interrupt essential processes during development and are thus normally detrimental or lethal. For example, if bacterial viruses are cultured under *permissive conditions* of 25°C, the mutant gene product is functional, infection proceeds normally, and new viruses are produced and can be studied. However, if bacterial viruses carrying temperature-sensitive mutations infect bacteria cultured at 42°C—the *restrictive condition*—infection progresses up to the point where the essential gene product is required (e.g., for viral assembly) and then arrests. Temperature-sensitive mutations are easily induced and isolated in viruses, and have added immensely to the study of viral genetics.

Nutritional Effects

Another category of phenotypes that are not always a direct reflection of the organism's genotype consists of **nutritional mutations.** In microorganisms, mutations that prevent synthesis of nutrient molecules are quite common, such as when an enzyme essential to a biosynthetic pathway becomes inactive. A microorganism bearing such a mutation is called an **auxotroph.** If the end product of a biochemical pathway can no longer be synthesized, and if that molecule is essential to normal growth and development, the mutation prevents growth and may be lethal. For example, if the bread mold *Neurospora* can no longer synthesize the amino acid leucine, proteins cannot be synthesized. If leucine is present in the growth medium, the detrimental effect is overcome. Nutritional mutants have been crucial to genetic studies in bacteria and also served as the basis for George Beadle and Edward Tatum's proposal, in the early 1940s, that one gene functions to produce one enzyme. (See Chapter 14.)

A slightly different set of circumstances exists in humans. The ingestion of certain dietary substances that normal individuals may consume without harm can adversely affect individuals with abnormal genetic constitutions. Often, a mutation may prevent an individual from metabolizing some substance commonly found in normal diets. For example, those afflicted with the genetic disorder **phenylketonuria (PKU)** cannot metabolize the amino acid phenylalanine. Those with **galactosemia** cannot metabolize galactose. Those with **lactose intolerance** cannot metabolize lactose. However, if the dietary intake of the involved molecule is drastically reduced or eliminated, the associated phenotype may be ameliorated.

Onset of Genetic Expression

Not all genetic traits become apparent at the same time during an organism's life span. In most cases, the age at which a mutant gene exerts a noticeable phenotype depends on events during the normal sequence of growth and development. In humans, the prenatal, infant, preadult, and adult phases require different genetic information. As a result, many severe inherited disorders are not manifested until after birth. Newborns with *Tay-Sachs disease,* discussed earlier in this chapter, appear to be phenotypically normal for the first few months. Then, developmental retardation, paralysis, and blindness ensue, and most affected children die around the age of 3.

The **Lesch–Nyhan syndrome,** inherited as an X-linked recessive disease, is characterized by abnormal nucleic acid metabolism (inability to salvage nitrogenous purine bases), leading to the accumulation of uric acid in blood and tissues, mental retardation, palsy, and self-mutilation of the lips and fingers. The disorder is due to a mutation in the gene encoding hypoxanthine-guanine phosphoribosyl transferase (HPRT). Newborns are normal for six to eight months prior to the onset of the first symptoms.

Still another example is **Duchenne muscular dystrophy (DMD),** an X-linked recessive disorder associated with progressive muscular wasting. It is not usually diagnosed until a child is 3 to 5 years old. Even with modern medical intervention, the disease is often fatal in the early 20s.

Perhaps the most delayed and highly variable age of onset for a genetic disorder in humans is seen in **Huntington**

disease. Inherited as an autosomal dominant disorder, Huntington disease affects the frontal lobes of the cerebral cortex, where progressive cell death occurs over a period of more than a decade. Brain deterioration is accompanied by spastic uncontrolled movements, intellectual deterioration, and ultimately death. While onset of these symptoms has been reported at all ages, they are most often initially observed between ages 30 and 50, with a mean onset age of 38 years.

These examples support the concept that gene products may play more essential roles at certain times during the life cycle of an organism. One may be able to tolerate the impact of a mutant gene for a considerable period of time without noticeable effect. At some point, however, a mutant phenotype is manifested. Perhaps this is the result of the internal physiological environment of an organism changing during development and with age.

Genetic Anticipation

Interest in studying the genetic onset of phenotypic expression has intensified with the discovery of heritable disorders that *exhibit a progressively earlier age of onset and an increased severity of the disorder in each successive generation.* This phenomenon is referred to as **genetic anticipation.**

Myotonic dystrophy (DM1), the most common type of adult muscular dystrophy, clearly illustrates genetic anticipation. Individuals afflicted with this autosomal dominant disorder exhibit extreme variation in the severity of symptoms. Mildly affected individuals develop cataracts as adults, but have little or no muscular weakness. Severely affected individuals demonstrate more extensive weakness, as well as myotonia (muscle hyperexcitability) and in some cases mental retardation. In its most extreme form, the disease is fatal just after birth. A great deal of excitement was generated in 1989, when C. J. Howeler and colleagues confirmed the correlation of increased severity and earlier onset with successive generations of inheritance. The researchers studied 61 parent–child pairs, and in 60 of the cases, age of onset was earlier and more severe in the child than in his or her affected parent.

In 1992, an explanation was put forward to explain both the molecular cause of the mutation responsible for DM1 and the basis of genetic anticipation in the disorder. As we will see later in the text, a three nucleotide sequence of DNA within the *DMPK* gene is repeated a variable number of times and is unstable (see Chapter 17). Normal individuals have about 5 to 35 copies of this sequence; affected individuals have between 150 and 2000 copies. Those with a greater number of repeats are more severely affected. The most remarkable observation was that, in successive generations of DM1 individuals, the size of the repeated segment increases. We now know that the RNA transcribed from mutant genes is the culprit in the disorder and alters the expression of still other genes. Several other inherited human disorders, including a second form of myotonic dystrophy (DM2), fragile-X syndrome, Kennedy disease, and Huntington disease, also reveal an association between the number of copies of a repeated DNA sequence and disease severity.

GENETICS, ETHICS, AND SOCIETY

Nature versus Nurture: Is the Debate Over?

Since ancient times, philosophers and scientists have debated the question of what makes us human. Are we creatures of our inheritance, or are we shaped by our environmental and social influences? That is, are we the product of nature or nurture? The answer to this question has changed significantly over the centuries and continues to evolve to this day.

From the late 1600s until the 1960s, the predominant view was that human traits are shaped overwhelmingly by our social and cultural environments—the *blank-slate* hypothesis. Even the influence of Darwin's theory of evolution, which proposed an almost-exclusive role for heredity in development (known as *genetic determinism*) was not able to displace this hypothesis.

The blank-slate hypothesis was more seriously challenged in the 1970s and 1980s when genetic studies in animals and twin studies in humans revealed a key role for genes in determining behavioral and intellectual traits. At this point, a backlash occurred, with proponents of the blank-slate hypothesis arguing that proponents of genetic determinism were engaging in political and ethical incorrectness that encouraged the idea of innate class distinctions.

By the 1990s, as genetics and genomic sequencing studies flourished, the debate shifted again. Currently, it is proposed that both genetics and environment together shape our traits. Although many studies show that nature and nurture both contribute, there is considerable variation in the ratios of each that determine each physical trait, disease susceptibility, and behavioral characteristic. For example, the risk of developing bipolar disorder is given as 69 percent due to genetics and 32 percent due to environmental factors. However, the risk of developing an eating disorder is approximately 40 percent due to genetics and 60 percent due to environmental factors. Every trait examined, even clear-cut human genetic disorders, has at least some genetic and some environmental contribution.

Many developmental and social scientists now declare that the nature–nurture debate is over and should be abandoned. They argue that the answer is clear and the assumed conflict between genes and the environment is nonexistent.

But is it?

(continued)

Genetics, Ethics, and Society, continued

Your Turn

Take time, individually or in groups, to discuss the following questions. Investigate the references and links dealing with the ethical and technical challenges surrounding the nature–nurture debate.

1. Given that it is unethical to conduct controlled experiments on humans to determine the relative roles of genes and the environment, researchers have resorted to various indirect methods, including twin studies. Describe several types of twin studies, what information can be extracted from them, and their shortcomings.

   An overview of twin studies can be found at (https://www.revolvy.com/main/index.php?s=Twin%20study&item_type=topic).

2. Supporters of genetic determinism point out that many single-gene Mendelian disorders such as Huntington disease, Tay-Sachs disease, phenylketonuria, and sickle-cell anemia have clear genetic causations. However, even these conditions can be influenced by environmental factors. What are some of these environmental influences? Explain how they do, or do not, affect the nature–nurture debate.

   For an introduction to this topic, see Collins, F. S., et al. (2001). Heredity and humanity. *New Republic* (http://www.arn.org/docs2/news/heredityandhumanity0711.htm).

3. The current widely accepted idea is that both nature and nurture influence all human traits, both physical and behavioral. Do you agree with these latest arguments, and why?

   For a description of a recent major study on these controversial topics, see Tan, M. (2015), Are we products of nature or nurture? Science answers age-old question, at (https://www.theguardian.com/science/2015/may/19/are-we-products-of-nature-or-nuture-science-answers-age-old-question). *Also see* Polderman, T. J., et al. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47:702–709.

CASE STUDY Should the child be deaf?

Researching their family histories, a deaf couple learns that some of their relatives back through several generations were also deaf. They plan to have a family and decide to discuss with a clinical geneticist the possibility that some or all of their children may be deaf. The geneticist informs them that without intervention they very likely will have a combination of deaf and normal hearing children, but that by using a combination of embryo testing and *in vitro* fertilization, it will be possible to select embryos with normal hearing for implantation. The couple opts for this method, and their first child has normal hearing.

1. Is it likely that these parents inherited their deafness as a recessive or dominant trait? What observations in the family histories would support either conclusion?
2. If the parents had asked to select embryos for deafness so that the child would share the characteristics and culture of his or her parents, would it be ethical for the physicians at the clinic to agree to test embryos for the presence of deafness and select these embryos for implantation?

See Draper, H., and Chadwick, R. (1999). Beware! Preimplantation genetic diagnosis may solve some old problems but it also raises new ones. *J. Med. Ethics* 25:114–120.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Since Mendel's work was rediscovered, transmission genetics has been expanded to include many alternative modes of inheritance, including the study of incomplete dominance, codominance, multiple alleles, and lethal alleles.
2. Mendel's classic F_2 ratio is often modified in instances when gene interaction controls phenotypic variation. Many such instances involve epistasis, whereby the expression of one gene influences or inhibits the expression of another gene.
3. Complementation analysis determines whether independently isolated mutations that produce similar phenotypes are alleles of one another, or whether they represent separate genes.
4. Pleiotropy refers to multiple phenotypic effects caused by a single mutation.
5. Genes located on the X chromosome result in a characteristic mode of genetic transmission referred to as X-linkage, displaying so-called crisscross inheritance, whereby affected mothers pass X-linked traits to all of their sons.
6. Sex-limited and sex-influenced inheritance occurs when the sex of the organism affects the phenotype controlled by a gene located on an autosome.
7. Phenotypic expression is not always the direct reflection of the genotype. Variable expressivity may be observed, or a percentage of organisms may not express the expected phenotype at all, the basis of the penetrance of a mutant allele. In addition, the phenotype can be modified by genetic background, temperature, and nutrition. Finally, the onset of expression of a gene may vary during the lifetime of an organism, and even become progressively more severe in subsequent generations (genetic anticipation).

INSIGHTS AND SOLUTIONS

Genetic problems take on added complexity if they involve two independent characters and multiple alleles, incomplete dominance, or epistasis. The most difficult types of problems are those that pioneering geneticists faced during laboratory or field studies. They had to determine the mode of inheritance by working backward from the observations of offspring to parents of unknown genotype.

1. Consider the problem of comb-shape inheritance in chickens, where walnut, rose, pea, and single are observed as distinct phenotypes. These variations are shown in the accompanying photographs. Considering the following data, determine how comb shape is inherited and what genotypes are present in the P1 generation of each cross.

Cross 1:	single × single	⟶	all single
Cross 2:	walnut × walnut	⟶	all walnut
Cross 3:	rose × pea	⟶	all walnut
Cross 4:	F_1 × F_1 of cross 3		
	walnut × walnut	⟶	93 walnut 28 rose 32 pea 10 single

Solution: At first glance, this problem appears quite difficult. However, working systematically and breaking the analysis into steps simplify it. To start, look at the data carefully for any useful information. Once you identify something that is clearly helpful, follow an empirical approach; that is, formulate a hypothesis and test it against the given data. Look for a pattern of inheritance that is consistent with all cases.

This problem gives two immediately useful facts. First, in cross 1, P_1 singles breed true. Second, while P_1 walnut breeds true in cross 2, a walnut phenotype is also produced in cross 3 between rose and pea. When these F_1 walnuts are mated in cross 4, all four comb shapes are produced in a ratio that approximates 9:3:3:1. This observation immediately suggests a cross involving two separate genes, because the resulting data display the same ratio as in Mendel's dihybrid crosses. Since only one character is involved (comb shape), epistasis may be occurring. This could serve as your working hypothesis, and you must now propose how the two gene pairs "interact" to produce each phenotype.

If you call the allele pairs *A*, *a* and *B*, *b*, you might predict that because walnut represents 9/16 of the offspring in cross 4, *A*–*B*– will produce walnut. (Recall that *A*– and *B*– mean *AA* or *Aa* and *BB* or *Bb*, respectively.) You might also hypothesize that in cross 2, the genotypes are *AABB* × *AABB* where walnut bred true.

The phenotype representing 1/16 of the offspring of cross 4 is single; therefore you could predict that the single phenotype is the result of the *aabb* genotype. This is consistent with cross 1.

Now you have only to determine the genotypes for rose and pea. The most logical prediction is that at least one dominant *A* or *B* allele combined with the double recessive condition of the other allele pair accounts for these phenotypes. For example,

A–*bb* ⟶ rose
aaB– ⟶ pea

If *AAbb* (rose) is crossed with *aaBB* (pea) in cross 3, all offspring would be *AaBb* (walnut). This is consistent with the data, and you need now look at only cross 4. We predict these walnut genotypes to be *AaBb*, and from the cross

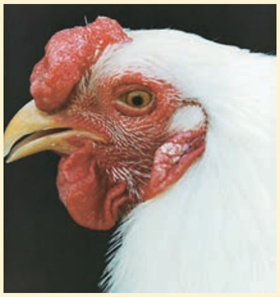
AaBb (walnut) × *AaBb* (walnut)

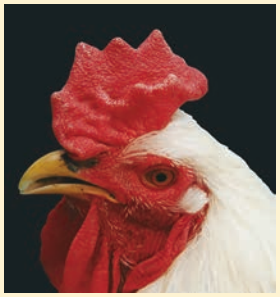
we expect

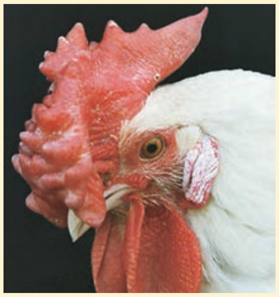
9/16	*A*–*B*–	(walnut)
3/16	*A*–*bb*	(rose)
3/16	*aaB*–	(pea)
1/16	*aabb*	(single)

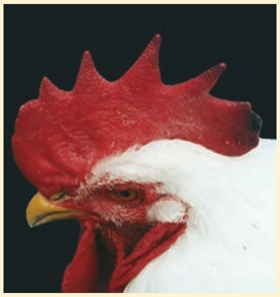
Our prediction is consistent with the data given. The initial hypothesis of the interaction of two gene pairs proves consistent throughout, and the problem is solved.

This problem demonstrates the usefulness of a basic theoretical knowledge of transmission genetics. With such knowledge, you can search for clues that will enable you to proceed in a stepwise fashion toward a solution. Mastering problem-solving requires practice, but can give you a great deal of satisfaction. Apply the same general approach to the following problems.


Walnut


Pea


Rose


Single

2. In radishes, flower color may be red, purple, or white. The edible portion of the radish may be long or oval. When only flower color is studied, no dominance is evident and red × white crosses yield all purple. If these F_1 purples are interbred, the F_2 generation consists of 1/4 red: 1/2 purple: 1/4 white. Regarding radish shape, long is dominant to oval in a normal Mendelian fashion.

(continued)

Insights and Solutions—continued

(a) Determine the F_1 and F_2 phenotypes from a cross between a true-breeding red, long radish and a radish that is white and oval. Be sure to define all gene symbols at the start.

(b) A red oval plant was crossed with a plant of unknown genotype and phenotype, yielding the following offspring:

103 red long: 101 red oval
98 purple long: 100 purple oval

Determine the genotype and phenotype of the unknown plant.

Solution: First, establish gene symbols:

RR = red	$O-$ = long
Rr = purple	oo = oval
rr = white	

(a) This is a modified dihybrid cross where the gene pair controlling color exhibits incomplete dominance. Shape is controlled conventionally.

P_1: *RROO* × *rroo*
(red long) (white oval)

F_1: all *RrOo* (purple long)

$F_1 \times F_1$: *RrOo* × *RrOo*

F_2:	1/4*RR*	3/4 *O*–	3/16 *RRO*–	red long
		1/4 *oo*	1/16 *RRoo*	red oval
	2/4*Rr*	3/4 *O*–	6/16 *RrO*–	purple long
		1/4 *oo*	2/16 *Rroo*	purple oval
	1/4*rr*	3/4 *O*–	3/16 *rrO*–	white long
		1/4 *oo*	1/16 *rroo*	white oval

Note that to generate the F_2 results, we have used the forked-line method. First, we consider the outcome of crossing F_1 parents for the color genes (*Rr* × *Rr*). Then the outcome of shape is considered (*Oo* × *Oo*).

(b) The two characters appear to be inherited independently, so consider them separately. The data indicate a 1/4:1/4:1/4:1/4 proportion. First, consider color:

P_1:	red × ???	(unknown)
F_1:	204 red	(1/2)
	198 purple	(1/2)

Because the red parent must be *RR*, the unknown must have a genotype of *Rr* to produce these results. Thus it is purple. Now, consider shape:

P_1:	oval × ???	(unknown)
F_1:	201 long	(1/2)
	201 oval	(1/2)

Since the oval plant must be *oo*, the unknown plant must have a genotype of *Oo* to produce these results. Thus it is long. The unknown plant is

RrOo purple long

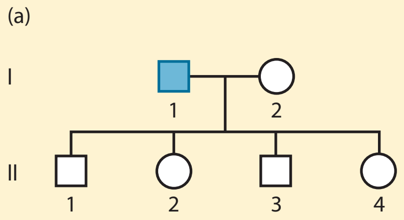
3. In humans, red–green color blindness is inherited as an X-linked recessive trait. A woman with normal vision whose father is color-blind marries a male who has normal vision. Predict the color vision of their male and female offspring.

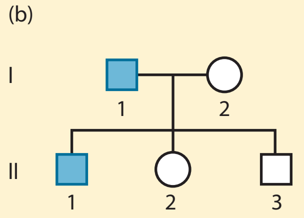
Solution: The female is heterozygous, since she inherited an X chromosome with the mutant allele from her father. Her husband is normal. Therefore, the parental genotypes are

Cc × *C*↾ (↾ represents the Y chromosome)

All female offspring are normal (*CC* or *Cc*). One-half of the male children will be color-blind (c↾), and the other half will have normal vision (*C*↾).

4. Consider the two very limited unrelated pedigrees shown here. Of the four combinations of X-linked recessive, X-linked dominant, autosomal recessive, and autosomal dominant, which modes of inheritance can be absolutely ruled out in each case?





Solution: For both pedigrees, X-linked recessive and autosomal recessive remain possible, provided that the maternal parent is heterozygous in pedigree (b). Autosomal dominance seems at first glance unlikely in pedigree (a), since at least half of the offspring should express a dominant trait expressed by one of their parents. However, while it is true that if the affected parent carries an autosomal dominant gene heterozygously, each offspring has a 50 percent chance of inheriting and expressing the mutant gene, the sample size of four offspring is too small to rule this possibility out. In pedigree (b), autosomal dominance is clearly possible. In both cases, one can rule out X-linked dominance because the female offspring would inherit and express the dominant allele, and they do not express the trait in either pedigree.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on extensions and modifications of Mendelian principles and ratios. In the process, we encountered many opportunities to consider how this information was acquired. On the basis of these discussions, what answers would you propose to the following fundamental questions?
(a) How were early geneticists able to ascertain inheritance patterns that did not fit typical Mendelian ratios?
(b) How did geneticists determine that inheritance of some phenotypic characteristics involves the interactions of two or more gene pairs? How were they able to determine how many gene pairs were involved?
(c) How do we know that specific genes are located on the sex-determining chromosomes rather than on autosomes?
(d) For genes whose expression seems to be tied to the sex of individuals, how do we know whether a gene is X-linked in contrast to exhibiting sex-limited or sex-influenced inheritance?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p. 62. These all relate to exceptions to the inheritance patterns encountered by Mendel. Write a short essay that explains why multiple and lethal alleles often result in a modification of the classic Mendelian monohybrid and dihybrid ratios.

3. In shorthorn cattle, coat color may be red, white, or roan. Roan is an intermediate phenotype expressed as a mixture of red and white hairs. The following data were obtained from various crosses:

red	×	red	⟶	all red
white	×	white	⟶	all white
red	×	white	⟶	all roan
roan	×	roan	⟶	1/4 red:1/2 roan: 1/4 white

How is coat color inherited? What are the genotypes of parents and offspring for each cross?

4. In foxes, two alleles of a single gene, *P* and *p*, may result in lethality (*PP*), platinum coat (*Pp*), or silver coat (*pp*). What ratio is obtained when platinum foxes are interbred? Is the *P* allele behaving dominantly or recessively in causing (a) lethality; (b) platinum coat color?

5. In mice, a short-tailed mutant was discovered. When it was crossed to a normal long-tailed mouse, 4 offspring were short-tailed and 3 were long-tailed. Two short-tailed mice from the F_1 generation were selected and crossed. They produced 6 short-tailed and 3 long-tailed mice. These genetic experiments were repeated three times with approximately the same results. What genetic ratios are illustrated? Hypothesize the mode of inheritance and diagram the crosses.

6. List all possible genotypes for the A, B, AB, and O phenotypes. Is the mode of inheritance of the ABO blood types representative of dominance, recessiveness, or codominance?

7. With regard to the ABO blood types in humans, determine the genotype of the male parent and female parent shown here:

Male parent: Blood type B; mother type O
Female parent: Blood type A; father type B

Predict the blood types of the offspring that this couple may have and the expected proportion of each.

8. In a disputed parentage case, the child is blood type O, while the mother is blood type A. What blood type would exclude a male from being the father? Would the other blood types prove that a particular male was the father?

9. The A and B antigens in humans may be found in water-soluble form in secretions, including saliva, of some individuals (*Se*/*Se* and *Se*/*se*) but not in others (*se*/*se*). The population thus contains “secretors” and “nonsecretors.”
(a) Determine the proportion of various phenotypes (blood type and ability to secrete) in matings between individuals that are blood type AB and type O, both of whom are *Se*/*se*.
(b) How will the results of such matings change if both parents are heterozygous for the gene controlling the synthesis of the H substance (*Hh*)?

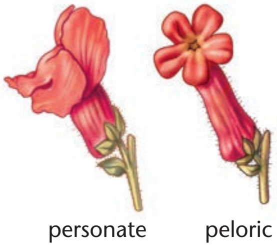
10. In chickens, a condition referred to as “creeper” exists whereby the bird has very short legs and wings and appears to be creeping when it walks. If creepers are bred to normal chickens, one-half of the offspring are normal and one-half are creepers. Creepers never breed true. If bred together, they yield two-thirds creepers and one-third normal. Propose an explanation for the inheritance of this condition.

11. In rabbits, a series of multiple alleles controls coat color in the following way: *C* is dominant to all other alleles and causes full color. The chinchilla phenotype is due to the c^{ch} allele, which is dominant to all alleles other than *C*. The c^h allele, dominant only to c^a (albino), results in the Himalayan coat color. Thus, the order of dominance is $C > c^{ch} > c^h > c^a$. For each of the following three cases, the phenotypes of the P_1 generations of two crosses are shown, as well as the phenotype of *one member* of the F_1 generation.

	P_1 Phenotypes		F_1 Phenotypes
	Himalayan × Himalayan	⟶	albino
(a)			× → ??
	full color × albino	⟶	chinchilla
	albino × chinchilla	⟶	albino
(b)			× → ??
	full color × albino	⟶	full color
	chinchilla × albino	⟶	Himalayan
(c)			× → ??
	full color × albino	⟶	Himalayan

For each case, determine the genotypes of the P_1 generation and the F_1 offspring, and predict the results of making each indicated cross between F_1 individuals.

12. Three gene pairs located on separate autosomes determine flower color and shape as well as plant height. The first pair exhibits incomplete dominance, where the color can be red, pink (the heterozygote), or white. The second pair leads to personate (dominant) or peloric (recessive) flower shape, while the third gene pair produces either the dominant tall trait or the recessive dwarf trait. Homozygous plants that are red, personate, and tall are crossed to those that are white, peloric, and dwarf. Determine the F_1 genotype(s) and phenotype(s). If the F_1 plants are interbred, what proportion of the offspring will exhibit the same phenotype as the F_1 plants?



13. As in Problem 12, flower color may be red, white, or pink, and flower shape may be personate or peloric. For the following crosses, determine the P_1 and F_1 genotypes:

(a) red, peloric × white, personate

↓

F_1: all pink, personate

(b) red, personate × white, peloric

↓

F_1: all pink, personate

$$\text{(c) pink, personate} \times \text{red, peloric} \rightarrow F_1 \begin{cases} 1/4 \text{ red, personate} \\ 1/4 \text{ red, peloric} \\ 1/4 \text{ pink, peloric} \\ 1/4 \text{ pink, personate} \end{cases}$$

$$\text{(d) pink, personate} \times \text{white, peloric} \rightarrow F_1 \begin{cases} 1/4 \text{ white, personate} \\ 1/4 \text{ white, peloric} \\ 1/4 \text{ pink, personate} \\ 1/4 \text{ pink, peloric} \end{cases}$$

(c) What phenotypic ratios would result from crossing the F_1 of (a) to the F_1 of (b)?

14. Horses can be cremello (a light cream color), chestnut (a brownish color), or palomino (a golden color with white in the horse's tail and mane). Of these phenotypes, only palominos never breed true.

cremello × palomino	⟶	1/2 cremello 1/2 palomino
chestnut × palomino	⟶	1/2 chestnut 1/2 palomino
palomino × palomino	⟶	1/4 chestnut 1/2 palomino 1/4 cremello

(a) From the results given above, determine the mode of inheritance by assigning gene symbols and indicating which genotypes yield which phenotypes.

(b) Predict the F_1 and F_2 results of many initial matings between cremello and chestnut horses.

Chestnut

Palomino

Cremello

15. With reference to the eye color phenotypes produced by the recessive, autosomal, unlinked *brown* and *scarlet* loci in *Drosophila* (see Figure 4.10), predict the F_1 and F_2 results of the following P_1 crosses. (Recall that when both the *brown* and *scarlet* alleles are homozygous, no pigment is produced, and the eyes are white.)

(a) wild type × white

(b) wild type × scarlet

(c) brown × white

16. Pigment in mouse fur is only produced when the *C* allele is present. Individuals of the *cc* genotype are white. If color is present, it may be determined by the *A*, *a* alleles. *AA* or *Aa* results in agouti color, while *aa* results in black coats.

(a) What F_1 and F_2 genotypic and phenotypic ratios are obtained from a cross between *AACC* and *aacc* mice?

(b) In three crosses between agouti females whose genotypes were unknown and males of the *aacc* genotype, the following phenotypic ratios were obtained:

(1)	8 agouti 8 white	(2)	9 agouti 10 black	(3)	4 agouti 5 black 10 white

What are the genotypes of these female parents?

17. In rats, the following genotypes of two independently assorting autosomal genes determine coat color:

A–B–	(gray)
A–bb	(yellow)
aaB–	(black)
aabb	(cream)

A third gene pair on a separate autosome determines whether or not any color will be produced. The *CC* and *Cc* genotypes allow color according to the expression of the *A* and *B* alleles. However, the *cc* genotype results in albino rats regardless of the *A* and *B* alleles present. Determine the F_1 phenotypic ratio of the following crosses:

(a) *AAbbCC* × *aaBBcc*

(b) *AaBBCC* × *AABbcc*

(c) *AaBbCc* × *AaBbcc*

(d) *AaBBCc* × *AaBBCc*

(e) *AABbCc* × *AABbcc*

18. Given the inheritance pattern of coat color in rats described in Problem 17, predict the genotype and phenotype of the parents who produced the following offspring:

(a) 9/16 gray: 3/16 yellow: 3/16 black: 1/16 cream

(b) 9/16 gray: 3/16 yellow: 4/16 albino

(c) 27/64 gray: 16/64 albino: 9/64 yellow: 9/64 black: 3/64 cream

(d) 3/8 black: 3/8 cream: 2/8 albino

(e) 3/8 black: 4/8 albino: 1/8 cream

19. In a species of the cat family, eye color can be gray, blue, green, or brown, and each trait is true breeding. In separate crosses involving homozygous parents, the following data were obtained:

Cross	P_1	F_1	F_2
A	green × gray	all green	3/4 green: 1/4 gray
B	green × brown	all green	3/4 green: 1/4 brown
C	gray × brown	all green	9/16 green: 3/16 brown 3/16 gray: 1/16 blue

(a) Analyze the data. How many genes are involved? Define gene symbols and indicate which genotypes yield each phenotype.

(b) In a cross between a gray-eyed cat and one of unknown genotype and phenotype, the F_1 generation was not observed. However, the F_2 resulted in the same F_2 ratio as in cross C. Determine the genotypes and phenotypes of the unknown P_1 and F_1 cats.

20. In a plant, a tall variety was crossed with a dwarf variety. All F_1 plants were tall. When $F_1 \times F_1$ plants were interbred, 9/16 of the F_2 were tall and 7/16 were dwarf.
(a) Explain the inheritance of height by indicating the number of gene pairs involved and by designating which genotypes yield tall and which yield dwarf. (Use dashes where appropriate.)
(b) What proportion of the F_2 plants will be true breeding if self-fertilized? List these genotypes.

21. In a unique species of plants, flowers may be yellow, blue, red, or mauve. All colors may be true breeding. If plants with blue flowers are crossed to red-flowered plants, all F_1 plants have yellow flowers. When these produced an F_2 generation, the following ratio was observed:

9/16 yellow: 3/16 blue: 3/16 red: 1/16 mauve

In still another cross using true-breeding parents, yellow-flowered plants are crossed with mauve-flowered plants. Again, all F_1 plants had yellow flowers and the F_2 showed a 9:3:3:1 ratio, as just shown.
(a) Describe the inheritance of flower color by defining gene symbols and designating which genotypes give rise to each of the four phenotypes.
(b) Determine the F_1 and F_2 results of a cross between true-breeding red and true-breeding mauve-flowered plants.

22. Five human matings (1–5), identified by both maternal and paternal phenotypes for ABO and MN blood-group antigen status, are shown on the left side of the following table:

	Parental Phenotypes				Offspring
(1)	A, M	×	A, N	(a)	A, N
(2)	B, M	×	B, M	(b)	O, N
(3)	O, N	×	B, N	(c)	O, MN
(4)	AB, M	×	O, N	(d)	B, M
(5)	AB, MN	×	AB, MN	(e)	B, MN

Each mating resulted in one of the five offspring shown in the right-hand column (a–e). Match each offspring with one correct set of parents, using each parental set only once. Is there more than one set of correct answers?

23. A husband and wife have normal vision, although both of their fathers are red–green color-blind, an inherited X-linked recessive condition. What is the probability that their first child will be (a) a normal son, (b) a normal daughter, (c) a color-blind son, (d) a color-blind daughter?

24. In humans, the ABO blood type is under the control of autosomal multiple alleles. Color blindness is a recessive X-linked trait. If two parents who are both type A and have normal vision produce a son who is color-blind and is type O, what is the probability that their next child will be a female who has normal vision and is type O?

25. In *Drosophila*, an X-linked recessive mutation, *scalloped* (*sd*), causes irregular wing margins. Diagram the F_1 and F_2 results if (a) a scalloped female is crossed with a normal male; (b) a scalloped male is crossed with a normal female. Compare these results with those that would be obtained if the *scalloped* gene were autosomal.

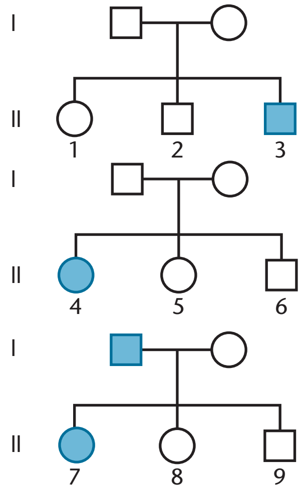
26. Another recessive mutation in *Drosophila*, *ebony* (*e*), is on an autosome (chromosome 3) and causes darkening of the body compared with wild-type flies. What phenotypic F_1 and F_2 male and female ratios will result if a scalloped-winged female with normal body color is crossed with a normal-winged ebony male? Work out this problem by both the Punnett square method and the forked-line method.

27. In *Drosophila*, the X-linked recessive mutation *vermilion* (*v*) causes bright red eyes, in contrast to the brick-red eyes of wild type. A separate autosomal recessive mutation, *suppressor of vermilion (su-v)*, causes flies homozygous or hemizygous for *v* to have wild-type eyes. In the absence of *vermilion* alleles, *su-v* has no effect on eye color. Determine the F_1 and F_2 phenotypic ratios from a cross between a female with wild-type alleles at the *vermilion* locus, but who is homozygous for *su-v*, with a *vermilion* male who has wild-type alleles at the *su-v* locus.

28. While *vermilion* is X-linked in *Drosophila* and causes the eye color to be bright red, *brown* is an autosomal recessive mutation that causes the eye to be brown. Flies carrying both mutations lose all pigmentation and are white-eyed. Predict the F_1 and F_2 results of the following crosses:
(a) vermilion females × brown males
(b) brown females × vermilion males
(c) white females × wild-type males

29. In a cross in *Drosophila* involving the X-linked recessive eye mutation *white* and the autosomally linked recessive eye mutation *sepia* (resulting in a dark eye), predict the F_1 and F_2 results of crossing true-breeding parents of the following phenotypes:
(a) white females × sepia males
(b) sepia females × white males
Note that *white* is epistatic to the expression of sepia.

30. Consider the three pedigrees below, all involving a single human trait.
(a) Which combination of conditions, if any, can be excluded?
dominant and X-linked
dominant and autosomal
recessive and X-linked
recessive and autosomal
(b) For each combination that you excluded, indicate the single individual in generation II (e.g., II-1, II-2) that was most instrumental in your decision to exclude it. If none were excluded, answer "none apply."



(c) Given your conclusions in part (a), indicate the genotype of the following individuals:

II-1, II-6, II-9

If more than one possibility applies, list all possibilities. Use the symbols *A* and *a* for the genotypes.

31. In goats, the development of the beard is due to a recessive gene. The following cross involving true-breeding goats was made and carried to the F_2 generation:

P_1: bearded female × beardless male
↓
F_1: all bearded males and beardless females

$$F_1 \times F_1 \rightarrow \begin{cases} 1/8 \text{ beardless males} \\ 3/8 \text{ bearded males} \\ 3/8 \text{ beardless females} \\ 1/8 \text{ bearded females} \end{cases}$$

Offer an explanation for the inheritance and expression of this trait, diagramming the cross. Propose one or more crosses to test your hypothesis.

32. Predict the F_1 and F_2 results of crossing a male fowl that is cock-feathered with a true-breeding hen-feathered female fowl. Recall that these traits are sex limited.

33. Two mothers give birth to sons at the same time at a busy urban hospital. The son of mother 1 is afflicted with hemophilia, a disease caused by an X-linked recessive allele. Neither parent has the disease. Mother 2 has a normal son, despite the fact that the father has hemophilia. Several years later, couple 1 sues the hospital, claiming that these two newborns were swapped in the nursery following their birth. As a genetic counselor, you are called to testify. What information can you provide the jury concerning the allegation?

34. Discuss the topic of phenotypic expression and the many factors that impinge on it.

35. Contrast penetrance and expressivity as the terms relate to phenotypic expression.

Extra-Spicy Problems

36. Labrador retrievers may be black, brown (chocolate), or golden (yellow) in color (see chapter-opening photo on p. 62). While each color may breed true, many different outcomes are seen when numerous litters are examined from a variety of matings where the parents are not necessarily true breeding. Following are just some of the many possibilities.

(a)	black	×	brown	⟶	all black
(b)	black	×	brown	⟶	1/2 black 1/2 brown
(c)	black	×	brown	⟶	3/4 black 1/4 golden
(d)	black	×	golden	⟶	all black
(e)	black	×	golden	⟶	4/8 golden 3/8 black 1/8 brown
(f)	black	×	golden	⟶	2/4 golden 1/4 black 1/4 brown
(g)	brown	×	brown	⟶	3/4 brown 1/4 golden
(h)	black	×	black	⟶	9/16 black 4/16 golden 3/16 brown

Propose a mode of inheritance that is consistent with these data, and indicate the corresponding genotypes of the parents in each mating. Indicate as well the genotypes of dogs that breed true for each color.

37. A true-breeding purple-leafed plant isolated from one side of El Yunque, the rain forest in Puerto Rico, was crossed to a true-breeding white variety found on the other side. The F_1 offspring were all purple. A large number of $F_1 \times F_1$ crosses produced the following results:

purple: 4219 white: 5781 (Total = 10,000)

Propose an explanation for the inheritance of leaf color. As a geneticist, how might you go about testing your hypothesis? Describe the genetic experiments that you would conduct.

38. In Dexter and Kerry cattle, animals may be polled (hornless) or horned. The Dexter animals have short legs, whereas the Kerry animals have long legs. When many offspring were obtained from matings between polled Kerrys and horned Dexters, half were found to be polled Dexters and half polled Kerrys. When these two types of F_1 cattle were mated to one another, the following F_2 data were obtained:

3/8 polled Dexters
3/8 polled Kerrys
1/8 horned Dexters
1/8 horned Kerrys

A geneticist was puzzled by these data and interviewed farmers who had bred these cattle for decades. She learned that Kerrys were true breeding. Dexters, on the other hand, were not true breeding and never produced as many offspring as Kerrys. Provide a genetic explanation for these observations.

39. A geneticist from an alien planet that prohibits genetic research brought with him to Earth two pure-breeding lines of frogs. One line croaks by *uttering* "rib-it rib-it" and has purple eyes. The other line croaks more softly by *muttering* "knee-deep knee-deep" and has green eyes. With a newfound freedom of inquiry, the geneticist mated the two types of frogs, producing F_1 frogs that were all utterers and had blue eyes. A large F_2 generation then yielded the following ratios:

27/64	blue-eyed, "rib-it" utterer
12/64	green-eyed, "rib-it" utterer
9/64	blue-eyed, "knee-deep" mutterer
9/64	purple-eyed, "rib-it" utterer
4/64	green-eyed, "knee-deep" mutterer
3/64	purple-eyed, "knee-deep" mutterer

(a) How many total gene pairs are involved in the inheritance of both traits? Support your answer.
(b) Of these, how many are controlling eye color? How can you tell? How many are controlling croaking?
(c) Assign gene symbols for all phenotypes and indicate the genotypes of the P_1 and F_1 frogs.
(d) Indicate the genotypes of the six F_2 phenotypes.

(e) After years of experiments, the geneticist isolated pure-breeding strains of all six F_2 phenotypes. Indicate the F_1 and F_2 phenotypic ratios of the following cross using these pure-breeding strains: blue-eyed, “knee-deep” mutterer × purple-eyed, “rib-it” utterer.

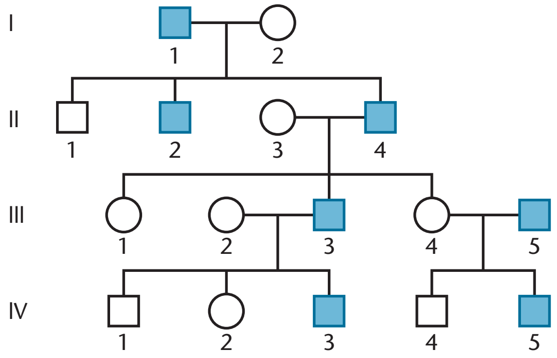
(f) One set of crosses with his true-breeding lines initially caused the geneticist some confusion. When he crossed true-breeding purple-eyed, “knee-deep” mutterers with true-breeding green-eyed, “knee-deep” mutterers, he often got different results. In some matings, all offspring were blue-eyed, “knee-deep” mutterers, but in other matings all offspring were purple-eyed, “knee-deep” mutterers. In still a third mating, 1/2 blue-eyed, “knee-deep” mutterers and 1/2 purple-eyed, “knee-deep” mutterers were observed. Explain why the results differed.

(g) In another experiment, the geneticist crossed two purple-eyed, “rib-it” *utterers* together with the results shown here:

9/16	purple-eyed, “rib-it” utterer
3/16	purple-eyed, “knee-deep” mutterer
3/16	green-eyed, “rib-it” utterer
1/16	green-eyed, “knee-deep” mutterer

What were the genotypes of the two parents?

40. The following pedigree is characteristic of an inherited condition known as male precocious puberty, where affected males show signs of puberty by age 4. Propose a genetic explanation of this phenotype.



41. Students taking a genetics exam were expected to answer the following question by converting data to a “meaningful ratio” and then solving the problem. The instructor assumed that the final ratio would reflect two gene pairs, and most correct answers did. Here is the exam question:

"Flowers may be white, orange, or brown. When plants with white flowers are crossed with plants with brown flowers, all the F_1 flowers are white. For F_2 flowers, the following data were obtained:

48	white
12	orange
4	brown

Convert the F_2 data to a meaningful ratio that allows you to explain the inheritance of color. Determine the number of genes involved and the genotypes that yield each phenotype."

(a) Solve the problem for two gene pairs. What is the final F_2 ratio?

(b) A number of students failed to reduce the ratio for two gene pairs as described above and solved the problem using three gene pairs. When examined carefully, their solution was deemed a valid response by the instructor. Solve the problem using three gene pairs.

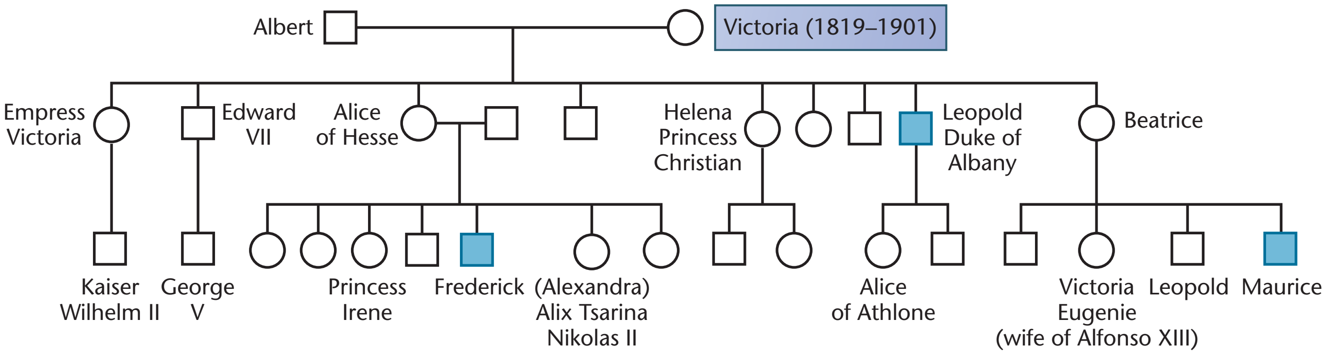
(c) We now have a dilemma. The data are consistent with two alternative mechanisms of inheritance. Propose an experiment that executes crosses involving the original parents that would distinguish between the two solutions proposed by the students. Explain how this experiment would resolve the dilemma.

42. In four o’clock plants, many flower colors are observed. In a cross involving two true-breeding strains, one crimson and the other white, all of the F_1 generation were rose color. In the F_2, four new phenotypes appeared along with the P_1 and F_1 parental colors. The following ratio was obtained:

1/16 crimson	4/16 rose
2/16 orange	2/16 pale yellow
1/16 yellow	4/16 white
2/16 magenta	

Propose an explanation for the inheritance of these flower colors.

43. Below is a partial pedigree of hemophilia in the British Royal Family descended from Queen Victoria, who is believed to be the original “carrier” in this pedigree. Analyze the pedigree and indicate which females are also certain to be carriers. What is the probability that Princess Irene is a carrier?



5

Chromosome Mapping in Eukaryotes

Chiasmata present between synapsed homologs during the first meiotic prophase.

CHAPTER CONCEPTS

- Chromosomes in eukaryotes contain large numbers of genes, whose locations are fixed along the length of the chromosomes.
- Unless separated by crossing over, alleles on the same chromosome segregate as a unit during gamete formation.
- Crossing over between homologs during meiosis creates recombinant gametes with different combinations of alleles that enhance genetic variation.
- Crossing over between homologs serves as the basis for the construction of chromosome maps. The greater the distance between two genes on a chromosome, the higher the frequency of crossing over is between them.
- While exchanges also occur between sister chromatids during mitosis, no new recombinant chromatids are created.

Walter Sutton, along with Theodor Boveri, was instrumental in uniting the fields of cytology and genetics. As early as 1903, Sutton pointed out the likelihood that there must be many more "unit factors" than chromosomes in most organisms. Soon thereafter, genetic studies with several organisms revealed that certain genes segregate as if they were somehow joined or linked together. Further investigations showed that such genes are part of the same chromosome, and they may indeed be transmitted as a single unit. We now know that most chromosomes contain a very large number of genes. Those that are part of the same chromosome are said to be *linked* and to demonstrate **linkage** in genetic crosses.

Because the chromosome, not the gene, is the unit of transmission during meiosis, linked genes are not free to undergo independent assortment. Instead, the alleles at all loci of one chromosome should, in theory, be transmitted as a unit during gamete formation. However, in many instances this does not occur. As we saw during the first meiotic prophase, when homologs are paired, or synapsed, a reciprocal exchange of chromosome segments may take place (Chapter 2). This **crossing over** results in the reshuffling, or **recombination,** of the alleles between homologs and always occurs during the tetrad stage.

Crossing over is currently viewed as an actual physical breaking and rejoining process that occurs during meiosis. You can see an example in the micrograph that opens this chapter. The exchange of chromosome segments provides an enormous potential for genetic variation in the gametes formed by any individual. This type of variation, in combination with that resulting from independent assortment, ensures that all offspring will contain a diverse mixture of maternal and paternal alleles.

The frequency of crossing over between any two loci on a single chromosome is proportional to the distance between them. Thus, depending on which loci are being considered, the percentage of recombinant gametes varies. This correlation allows us to construct **chromosome maps,** which indicate the relative locations of genes on the chromosomes.

In this chapter, we will discuss linkage, crossing over, and chromosome mapping in more detail. We will also consider a variety of other topics involving the exchange of genetic information, concluding the chapter with the rather intriguing observation that exchange occurs even between sister chromatids in mitosis prior to their separation in prophase, although no new allelic combinations are produced.

5.1 Genes Linked on the Same Chromosome Segregate Together

A simplified overview of the major theme of this chapter is given in **Figure 5.1**, which contrasts the meiotic consequences of (a) independent assortment, (b) linkage *without* crossing over, and (c) linkage *with* crossing over. In **Figure 5.1(a)** we see the results of independent assortment of two pairs of chromosomes, each containing one heterozygous gene pair. No linkage is exhibited. When these same two chromosomes are observed in a large number of meiotic events, they are seen to form four genetically different gametes in equal proportions, each containing a different combination of alleles of the two genes.

Now let's compare these results with what occurs if the same genes are linked on the same chromosome. If no crossing over occurs between the two genes [**Figure 5.1(b)**], only two genetically different kinds of gametes are formed. Each gamete receives the alleles present on one homolog or the other, which is transmitted intact as the result of segregation. This case illustrates *complete linkage,* which produces only **parental,** or **noncrossover, gametes.** The two parental gametes are formed in equal proportions. Though complete linkage between two genes seldom occurs, it is useful to consider the theoretical consequences of this concept.

Figure 5.1(c) shows the results when crossing over occurs between two linked genes. As you can see, this

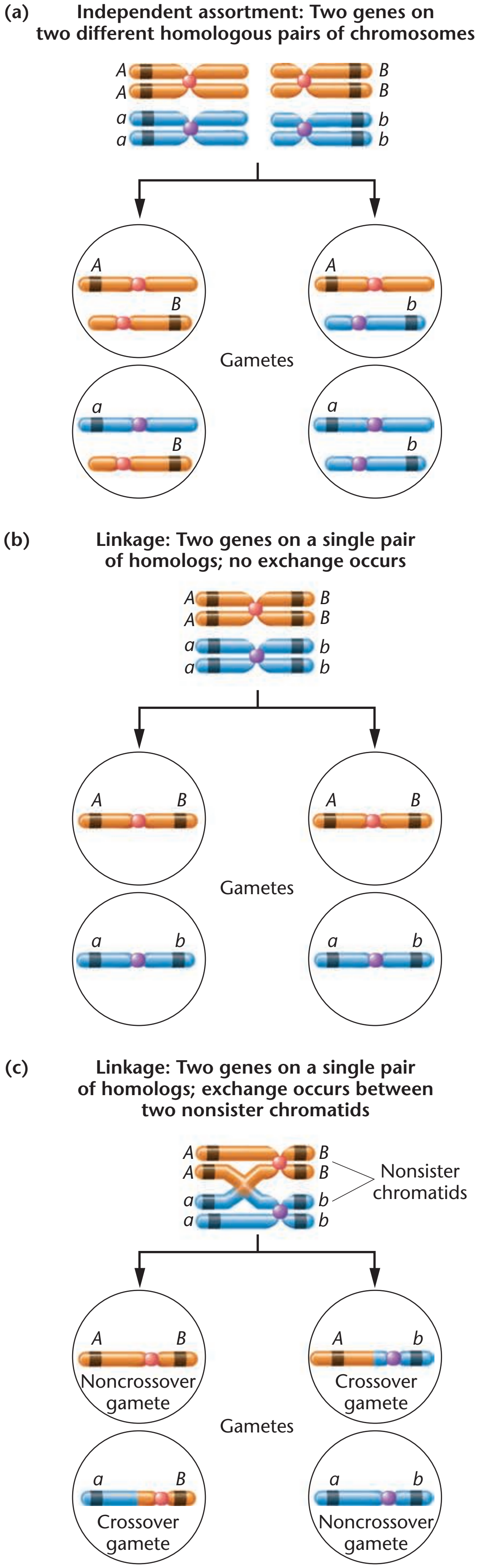

(a) Independent assortment: Two genes on two different homologous pairs of chromosomes
A
A
a
a
B
B
b
b
A
B
A
b
Gametes
a
B
a
b
(b) Linkage: Two genes on a single pair of homologs; no exchange occurs
A
A
a
a
B
B
b
b
A B
A B
Gametes
a b
a b
(c) Linkage: Two genes on a single pair of homologs; exchange occurs between two nonsister chromatids
A
A
a
a
B
B
b
b
Nonsister chromatids
A B
Noncrossover gamete
A b
Crossover gamete
Gametes
a B
Crossover gamete
a b
Noncrossover gamete

FIGURE 5.1 Results of gamete formation when two heterozygous genes are (a) on two different pairs of chromosomes; (b) on the same pair of homologs, but with no exchange occurring between them; and (c) on the same pair of homologs, but with an exchange occurring between two nonsister chromatids. Note in this and the following figures that members of homologous pairs of chromosomes are shown in two different colors. (This convention was established in Chapter 2; see, for example, Figure 2.7 and Figure 2.11.)

crossover involves only two nonsister chromatids of the four chromatids present in the tetrad. This exchange generates two new allele combinations, called **recombinant,** or **crossover, gametes.** The two chromatids not involved in the exchange result in noncrossover gametes, like those in Figure 5.1(b). Importantly, the frequency with which crossing over occurs between any two linked genes is proportional to the distance separating the respective loci along the chromosome. As the distance between the two genes increases, the proportion of recombinant gametes increases and that of the parental gametes decreases. In theory, two randomly selected genes can be so close to each other that crossover events are too infrequent to be easily detected. As shown in Figure 5.1(b), this complete linkage produces only parental gametes. On the other hand, if a small, but distinct, distance separates two genes, few recombinant and many parental gametes will be formed.

As we will discuss again later in this chapter, when the loci of two linked genes are far apart, the number of recombinant gametes approaches, but does not exceed, 50 percent. If 50 percent recombinants occur, the result is a 1 : 1 : 1 : 1 ratio of the four types (two parental and two recombinant gametes). In this case, transmission of two linked genes is indistinguishable from that of two unlinked, independently assorting genes. That is, the proportion of the four possible genotypes would be identical, as shown in Figure 5.1(a) and Figure 5.1(c).

The Linkage Ratio

If complete linkage exists between two genes because of their close proximity, and organisms heterozygous at both loci are mated, a unique F_2 phenotypic ratio results, which we designate the **linkage ratio.** To illustrate this ratio, let's consider a cross involving the closely linked, recessive, mutant genes *heavy* wing vein (*hv*) and *brown* eye (*bw*) in *Drosophila melanogaster* (**Figure 5.2**). The normal, wild-type alleles hv^+ and bw^+ are both dominant and result in thin wing veins and red eyes, respectively.

In this cross, flies with normal thin wing veins and mutant brown eyes are mated to flies with mutant heavy wing veins and normal red eyes. In more concise terms, heavy-veined flies are crossed with brown-eyed flies. Linked genes are represented by placing their allele designations (the genetic symbols established in Chapter 4) above and below a single or double horizontal line. Those above the line are located at loci on one homolog, and those below the line are located at the homologous loci on the other homolog. Thus, we represent the P_1 generation as follows:

$$P_1 : \frac{hv^+\,bw}{hv^+\,bw} \times \frac{hv\,bw^+}{hv\,bw^+}$$

thin, brown heavy, red

Because these genes are located on an autosome, no designation of male or female is necessary.

In the F_1 generation, each fly receives one chromosome of each pair from each parent. All flies are heterozygous for both gene pairs and exhibit the dominant traits of thin veins and red eyes:

$$F_1 = \frac{hv^+\,bw}{hv\ \ bw^+}$$

thin, red

As shown in **Figure 5.2(a)**, when the F_1 generation is interbred, each F_1 individual forms only parental gametes because of complete linkage. Following fertilization, the F_2 generation is produced in a 1 : 2 : 1 phenotypic and genotypic ratio. One-fourth of this generation shows thin wing veins and brown eyes; one-half shows both wild-type traits, namely, thin veins and red eyes; and one-fourth will show heavy wing veins and red eyes. Therefore, the ratio is 1 heavy : 2 wild : 1 brown. Such a 1 : 2 : 1 ratio is characteristic of complete linkage. Complete linkage is usually observed only when genes are very close together and the number of progeny is relatively small.

Figure 5.2(b) demonstrates the results of a testcross with the F_1 flies. Such a cross produces a 1 : 1 ratio of thin, brown and heavy, red flies. Had the genes controlling these traits been incompletely linked or located on separate autosomes, the testcross would have produced four phenotypes rather than two.

When large numbers of mutant genes in any given species are investigated, genes located on the same chromosome show evidence of linkage to one another. As a result, **linkage groups** can be identified, one for each chromosome. In theory, the number of linkage groups should correspond to the haploid number of chromosomes. In diploid organisms in which large numbers of mutant genes are available for genetic study, this correlation has been confirmed.

NOW SOLVE THIS

5.1 Consider two hypothetical recessive autosomal genes *a* and *b*, where a heterozygote is testcrossed to a double-homozygous mutant. Predict the phenotypic ratios under the following conditions:

(a) *a* and *b* are located on separate autosomes.

(b) *a* and *b* are linked on the same autosome but are so far apart that a crossover always occurs between them.

(c) *a* and *b* are linked on the same autosome but are so close together that a crossover almost never occurs.

■ **HINT:** *This problem involves an understanding of linkage, crossing over, and independent assortment. The key to its solution is to be aware that results are indistinguishable when two genes are unlinked compared to the case where they are linked but so far apart that crossing over always intervenes between them during meiosis.*

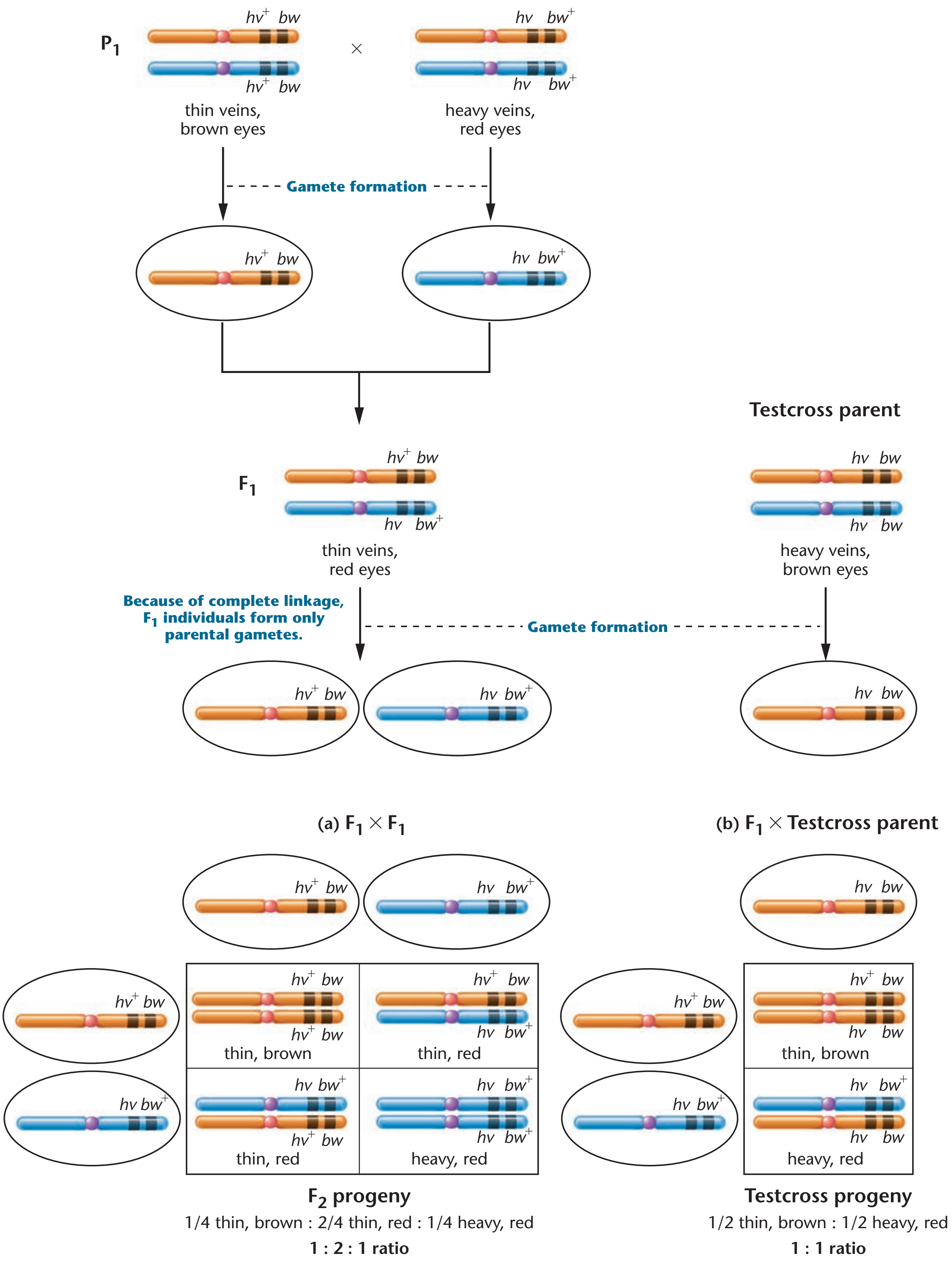


FIGURE 5.2 Results of a cross involving two genes located on the same chromosome and demonstrating complete linkage. (a) The F_2 results of the cross. (b) The results of a testcross involving the F_1 progeny.

5.2 Crossing Over Serves as the Basis for Determining the Distance between Genes in Chromosome Mapping

It is highly improbable that two randomly selected genes linked on the same chromosome will be so close to one another along the chromosome that they demonstrate complete linkage. Instead, crosses involving two such genes will almost always produce a percentage of offspring resulting from recombinant gametes. The percentage will vary depending on the distance between the two genes along the chromosome. This phenomenon was first explained in 1911 by two *Drosophila* geneticists, Thomas H. Morgan and his undergraduate student, Alfred H. Sturtevant.

Morgan and Crossing Over

As you may recall from our earlier discussion (see Chapter 4), Morgan was the first to discover the phenomenon of X-linkage. In his studies, he investigated numerous *Drosophila* mutations located on the X chromosome. His original analysis, based on crosses involving only one gene on the X chromosome, led to the discovery of X-linked inheritance. However, when he made crosses involving two X-linked genes, his results were initially puzzling. For example, female flies expressing the mutant *yellow* body (*y*) and *white* eyes (*w*) alleles were crossed with wild-type males (gray body and red eyes). The F_1 females were wild type, while the F_1 males expressed both mutant traits. In the F_2 the vast majority of the total offspring showed the expected parental phenotypes—yellow-bodied, white-eyed flies and wild-type flies (gray-bodied, red-eyed). The remaining flies, less than 1.0 percent, were either yellow-bodied with red eyes or gray-bodied with white eyes. It was as if the two mutant alleles had somehow separated from each other on the homolog during gamete formation in the F_1 female flies. This outcome is illustrated in cross A of **Figure 5.3**, using data later compiled by Sturtevant.

When Morgan studied other X-linked genes, the same basic pattern was observed, but the proportion of F_2 phenotypes differed. For example, when he crossed *white*-eye, *miniature*-wing mutants with wild-type flies, only 65.5 percent of all the F_2 flies showed the parental phenotypes, while 34.5 percent of the offspring appeared as if the mutant genes had been separated during gamete formation. This is illustrated in cross B of Figure 5.3, again using data subsequently compiled by Sturtevant.

Morgan was faced with two questions: (1) What was the source of gene separation and (2) why did the frequency of the apparent separation vary depending on the genes being studied? The answer Morgan proposed for the first question was based on his knowledge of earlier cytological observations made by F. A. Janssens and others. Janssens had observed that synapsed homologous chromosomes in meiosis wrapped around each other, creating **chiasmata** (sing. chiasma), X-shaped intersections where points of overlap are evident (see the photo on p. 94). Morgan proposed that these chiasmata could represent points of genetic exchange.

Regarding the crosses shown in Figure 5.3, Morgan postulated that if an exchange of chromosome material occurs during gamete formation, at a chiasma between the mutant genes on the two X chromosomes of the F_1 females, the unique phenotypes will occur. He suggested that such exchanges led to recombinant gametes in both the *yellow–white* cross and the *white–miniature* cross, as compared to the parental gametes that underwent no exchange. On the basis of this and other experimentation, Morgan concluded that linked genes are arranged in a linear sequence along the chromosome and that a variable frequency of exchange occurs between any two genes during gamete formation.

In answer to the second question, Morgan proposed that two genes located relatively close to each other along a chromosome are less likely to have a chiasma form between them than if the two genes are farther apart on the chromosome. Therefore, the closer two genes are, the less likely that a genetic exchange will occur between them. Morgan was the first to propose the term *crossing over* to describe the physical exchange leading to recombination.

Sturtevant and Mapping

Morgan's student, Alfred H. Sturtevant, was the first to realize that his mentor's proposal could be used to map the sequence of linked genes. According to Sturtevant,

> In a conversation with Morgan . . . I suddenly realized that the variations in strength of linkage, already attributed by Morgan to differences in the spatial separation of the genes, offered the possibility of determining sequences in the linear dimension of a chromosome. I went home and spent most of the night (to the neglect of my undergraduate homework) in producing the first chromosomal map.

Sturtevant, in a paper published in 1913, compiled data from numerous crosses made by Morgan and other geneticists involving recombination between the genes represented by the *yellow, white,* and *miniature* mutants. A subset of these data is shown in Figure 5.3. The frequencies of recombination between each pair of these three genes are as follows:

(1)	*yellow, white*	0.5%
(2)	*white, miniature*	34.5%
(3)	*yellow, miniature*	35.4%

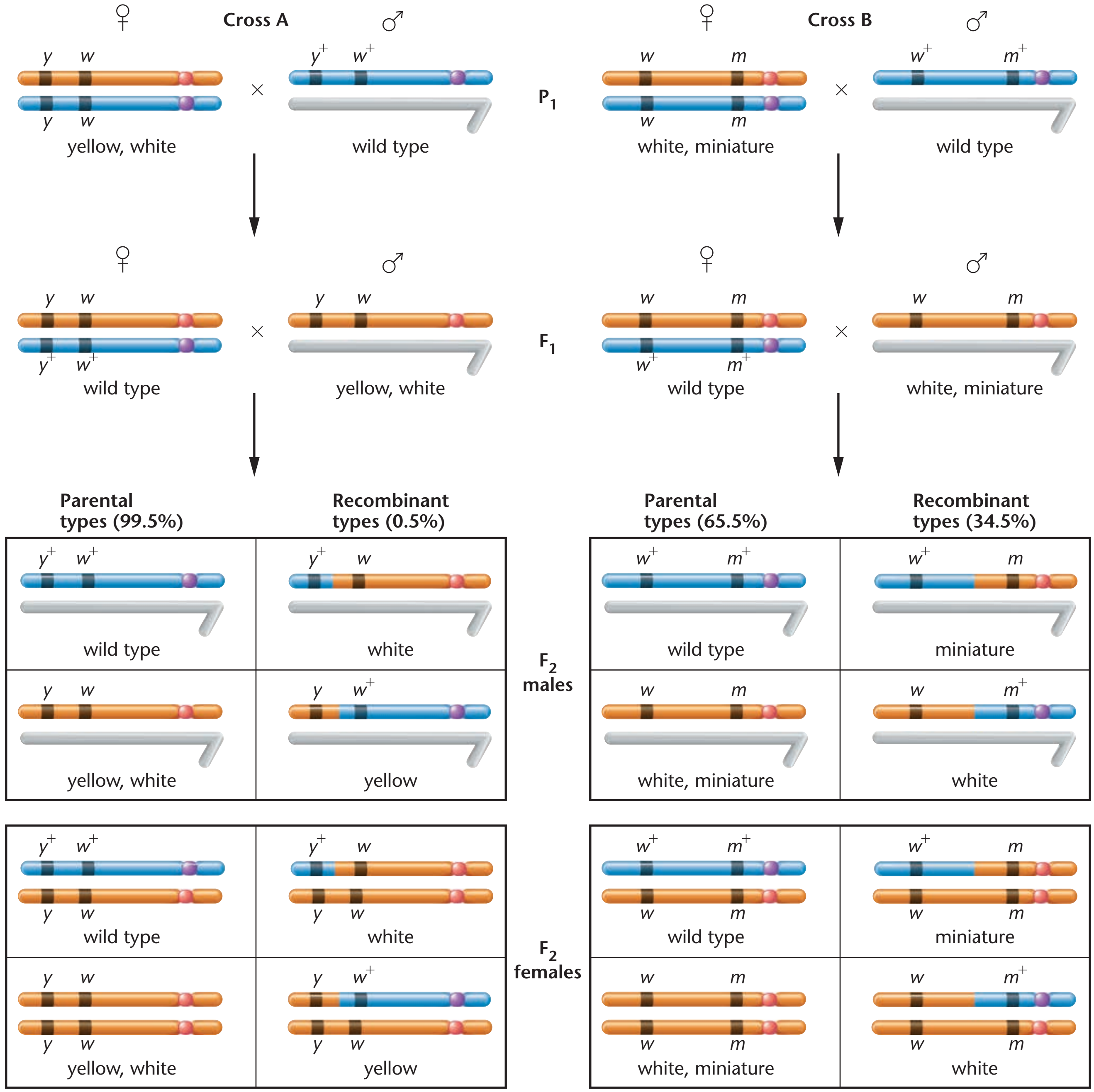


FIGURE 5.3 The F$_1$ and F$_2$ results of crosses involving the *yellow* (*y*), *white* (*w*) mutations (cross A), and the *white*, *miniature* (*m*) mutations (cross B), as compiled by Sturtevant. In cross A, 0.5 percent of the F$_2$ flies (males and females) demonstrate recombinant phenotypes, which express either *white* or *yellow*. In cross B, 34.5 percent of the F$_2$ flies (males and females) demonstrate recombinant phenotypes, which are either *miniature* or *white* mutants.

Because the sum of (1) and (2) approximately equals (3), Sturtevant suggested that the recombination frequencies between linked genes are additive. On this basis, he predicted that the order of the genes on the X chromosome is *yellow–white–miniature.* In arriving at this conclusion, he reasoned as follows: The *yellow* and *white* genes are apparently close to each other because the recombination frequency is low. However, both of these genes are quite far from the *miniature* gene because the *white–miniature* and *yellow–miniature* combinations show larger recombination frequencies. Because *miniature* shows more recombination with *yellow* than with *white* (35.4 percent vs. 34.5 percent), it follows that *white* is located between the other two genes, not outside of them.

Sturtevant knew from Morgan's work that the frequency of exchange could be used as an estimate of the distance between two genes or loci along the chromosome. He constructed a chromosome map of the three genes on the X chromosome, setting one map unit (mu) equal to 1 percent

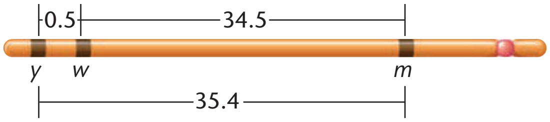


FIGURE 5.4 A map of the *yellow* (*y*), *white* (*w*), and *miniature* (*m*) genes on the X chromosome of *Drosophila melanogaster.* Each number represents the percentage of recombinant offspring produced in one of three crosses, each involving two different genes.

recombination between two genes.* The distance between *yellow* and *white* is thus 0.5 mu, and the distance between *yellow* and *miniature* is 35.4 mu. It follows that the distance between *white* and *miniature* should be 35.4 − 0.5 = 34.9 mu. This estimate is close to the actual frequency of recombination between *white* and *miniature* (34.5 mu). The map for these three genes is shown in Figure 5.4. The fact that these numbers do not add up perfectly is due to normal variation that one would expect between crosses, leading to the minor imprecisions encountered in independently conducted mapping experiments.

In addition to these three genes, Sturtevant considered crosses involving two other genes on the X chromosome and produced a more extensive map that included all five genes. He and a colleague, Calvin Bridges, soon began a search for autosomal linkage in *Drosophila*. By 1923, they had clearly shown that linkage and crossing over are not restricted to X-linked genes but could also be demonstrated with autosomes. During this work, they made another interesting observation. In *Drosophila,* crossing over was shown to occur only in females. The fact that no crossing over occurs in males made genetic mapping much less complex to analyze in *Drosophila*. While crossing over does occur in both sexes in most other organisms, crossing over in males is often observed to occur less frequently than in females. For example, in humans, such recombination occurs only about 60 percent as often in males compared to females.

Although many refinements have been added to chromosome mapping since Sturtevant's initial work, his basic principles are accepted as correct. These principles are used to produce detailed chromosome maps of organisms for which large numbers of linked mutant genes are known. Sturtevant's findings are also historically significant to the broader field of genetics. In 1910, the chromosomal theory of inheritance was still widely disputed—even Morgan was skeptical of this theory before he conducted his experiments. Research has now firmly established that chromosomes contain genes in a linear order and that these genes are the equivalent of Mendel's unit factors.

Single Crossovers

Why should the relative distance between two loci influence the amount of crossing over and recombination observed between them? During meiosis, a limited number of crossover events occur in each tetrad. These recombinant events occur randomly along the length of the tetrad. Therefore, the closer that two loci reside along the axis of the chromosome, the less likely that any **single crossover** event will occur between them. The same reasoning suggests that the farther apart two linked loci, the more likely a random crossover event will occur in between them.

In Figure 5.5(a), a single crossover occurs between two nonsister chromatids, but not between the two loci being studied; therefore, the crossover is undetected because no recombinant gametes are produced for the two traits of interest. In Figure 5.5(b), where the two loci under study are quite far apart, the crossover does occur between them, yielding gametes in which the traits of interest are recombined.

When a single crossover occurs between two nonsister chromatids, the other two chromatids of the tetrad are not involved in the exchange and enter the gamete unchanged. Even if a single crossover occurs 100 percent

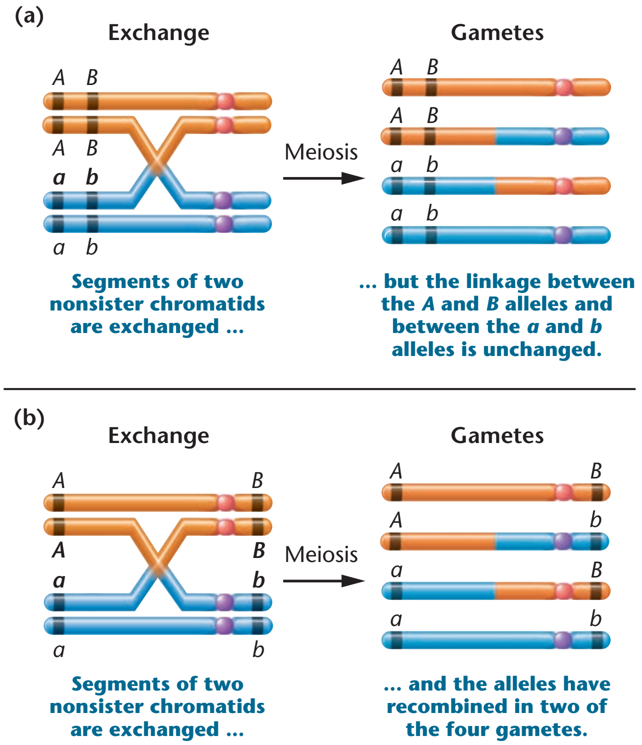

(a)
Exchange
Gametes
A B
A B
a b
a b
Meiosis
A B
A B
a b
a b
Segments of two nonsister chromatids are exchanged ...
... but the linkage between the A and B alleles and between the a and b alleles is unchanged.
(b)
Exchange
Gametes
A B
A B
a b
a b
Meiosis
A B
A b
a B
a b
Segments of two nonsister chromatids are exchanged ...
... and the alleles have recombined in two of the four gametes.

FIGURE 5.5 Two examples of a single crossover between two nonsister chromatids and the gametes subsequently produced. In (a) the exchange does not alter the linkage arrangement between the alleles of the two genes, only parental gametes are formed, and the exchange goes undetected. In (b) the exchange separates the alleles, resulting in recombinant gametes, which are detectable.

*In honor of Morgan's work, map units are often referred to as centi-Morgans (cM).

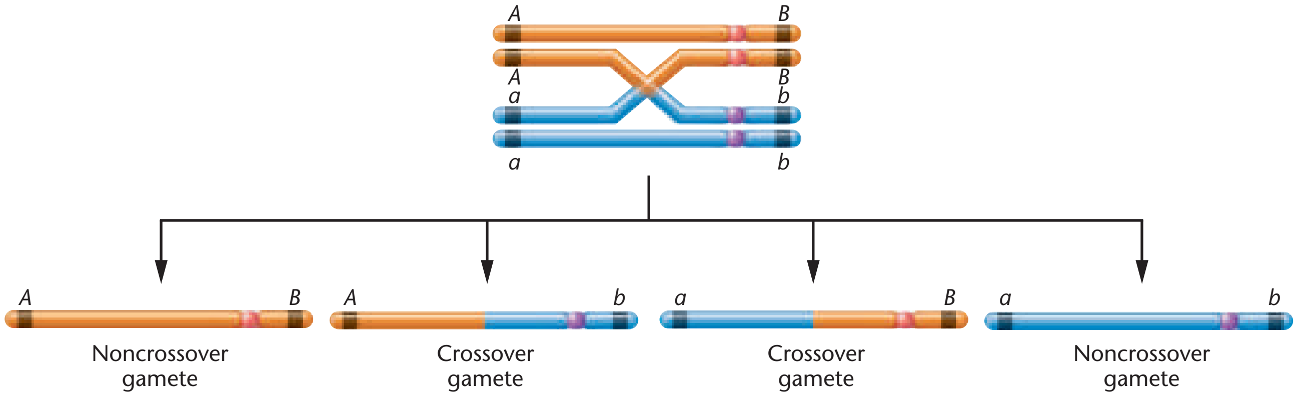


FIGURE 5.6 The consequences of a single exchange between two nonsister chromatids occurring in the tetrad stage. Two noncrossover (parental) and two crossover (recombinant) gametes are produced.

of the time between two linked genes, recombination is subsequently observed in only 50 percent of the potential gametes formed. This concept is diagrammed in Figure 5.6. Theoretically, if we assume only single exchanges between a given pair of loci and observe 20 percent recombinant gametes, we will conclude that crossing over actually occurs between these two loci in 40 percent of the tetrads. The general rule is that, under these conditions, the percentage of tetrads involved in an exchange between two genes is twice as great as the percentage of recombinant gametes produced. Therefore, the theoretical limit of observed recombination due to crossing over is 50 percent.

When two linked genes are more than 50 map units apart, a crossover can theoretically be expected to occur between them in 100 percent of the tetrads. If this prediction were achieved, each tetrad would yield equal proportions of the four gametes shown in Figure 5.6, just as if the genes were on different chromosomes and assorting independently. For a variety of reasons, this theoretical limit is seldom achieved.

5.3 Determining the Gene Sequence during Mapping Requires the Analysis of Multiple Crossovers

The study of single crossovers between two linked genes provides a basis for determining the *distance* between them. However, when many linked genes are studied, their *sequence* along the chromosome is more difficult to determine. Fortunately, the discovery that multiple crossovers occur between the chromatids of a tetrad has facilitated the process of producing more extensive chromosome maps. As we shall see next, when three or more linked genes are investigated simultaneously, it is possible to determine first the sequence of genes and then the distances between them.

NOW SOLVE THIS

5.2 With two pairs of genes involved (*P*/*p* and *Z*/*z*), a testcross (*ppzz*) with an organism of unknown genotype indicated that the gametes produced were in the following proportions:

PZ, 42.4%; *Pz*, 6.9%; *pZ*, 7.1%; *pz*, 43.6%

Draw all possible conclusions from these data.

■ **HINT:** *This problem involves an understanding of the proportionality between crossover frequency and distance between genes. The key to its solution is to be aware that noncrossover and crossover gametes occur in reciprocal pairs of approximately equal proportions.*

Multiple Exchanges

It is possible that in a single tetrad, two, three, or more exchanges will occur between nonsister chromatids as a result of several crossing over events. Double exchanges of genetic material result from **double crossovers (DCOs),** as shown in Figure 5.7. To study a double exchange, three gene pairs must be investigated, each heterozygous for two alleles. Before we determine the frequency of recombination among all three loci, let's review some simple probability calculations.

As we have seen, the probability of a single exchange occurring in between the *A* and *B* or the *B* and *C* genes is related directly to the distance between the respective loci. The closer *A* is to *B* and *B* is to *C,* the less likely it is that a single exchange will occur in between either of the two sets of loci. In the case of a double crossover, two separate and independent events or exchanges must occur simultaneously. The mathematical probability of two independent events occurring simultaneously is equal to the product of

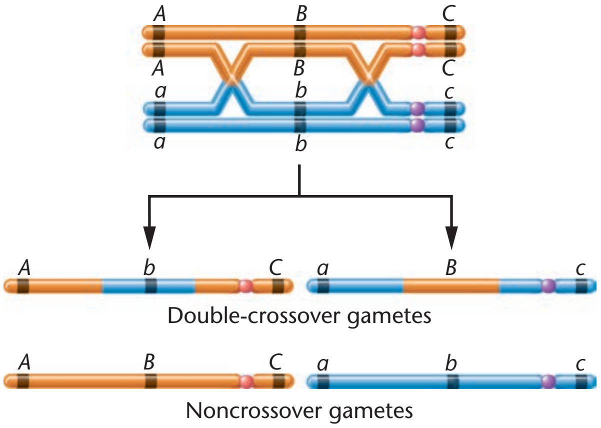


FIGURE 5.7 Consequences of a double exchange occurring between two nonsister chromatids. Because the exchanges involve only two chromatids, two noncrossover gametes and two double-crossover gametes are produced. The chapter-opening photograph on p. 94 illustrates two chiasmata present in a tetrad isolated during the first meiotic prophase stage.

the individual probabilities. (This is the *product law* introduced in Chapter 3.)

Suppose that crossover gametes resulting from single exchanges are recovered 20 percent of the time ($p = 0.20$) between *A* and *B*, and 30 percent of the time ($p = 0.30$) between *B* and *C*. The probability of recovering a double-crossover gamete arising from two exchanges (between *A* and *B* and between *B* and *C*) is predicted to be $(0.20)(0.30) = 0.06$, or 6 percent. It is apparent from this calculation that the expected frequency of double-crossover gametes is always expected to be much lower than that of either single-crossover class of gametes.

If three genes are relatively close together along one chromosome, the expected frequency of double-crossover gametes is extremely low. For example, suppose that the *A*–*B* distance in Figure 5.7 is 3 mu and the *B*–*C* distance is 2 mu. The expected double-crossover frequency is $(0.03)(0.02) = 0.0006$, or 0.06 percent. This translates to only six events in 10,000. Thus in a mapping experiment where closely linked genes are involved, very large numbers of offspring are required to detect double-crossover events. In this example, it is unlikely that a double crossover will be observed even if 1000 offspring are examined. Thus, it is evident that if four or five genes are being mapped, even fewer triple and quadruple crossovers can be expected to occur.

Three-Point Mapping in *Drosophila*

The information presented in Section 5.2 enables us to map three or more linked genes in a single cross. To illustrate the mapping process in its entirety, we examine two situations involving three linked genes in two quite different organisms.

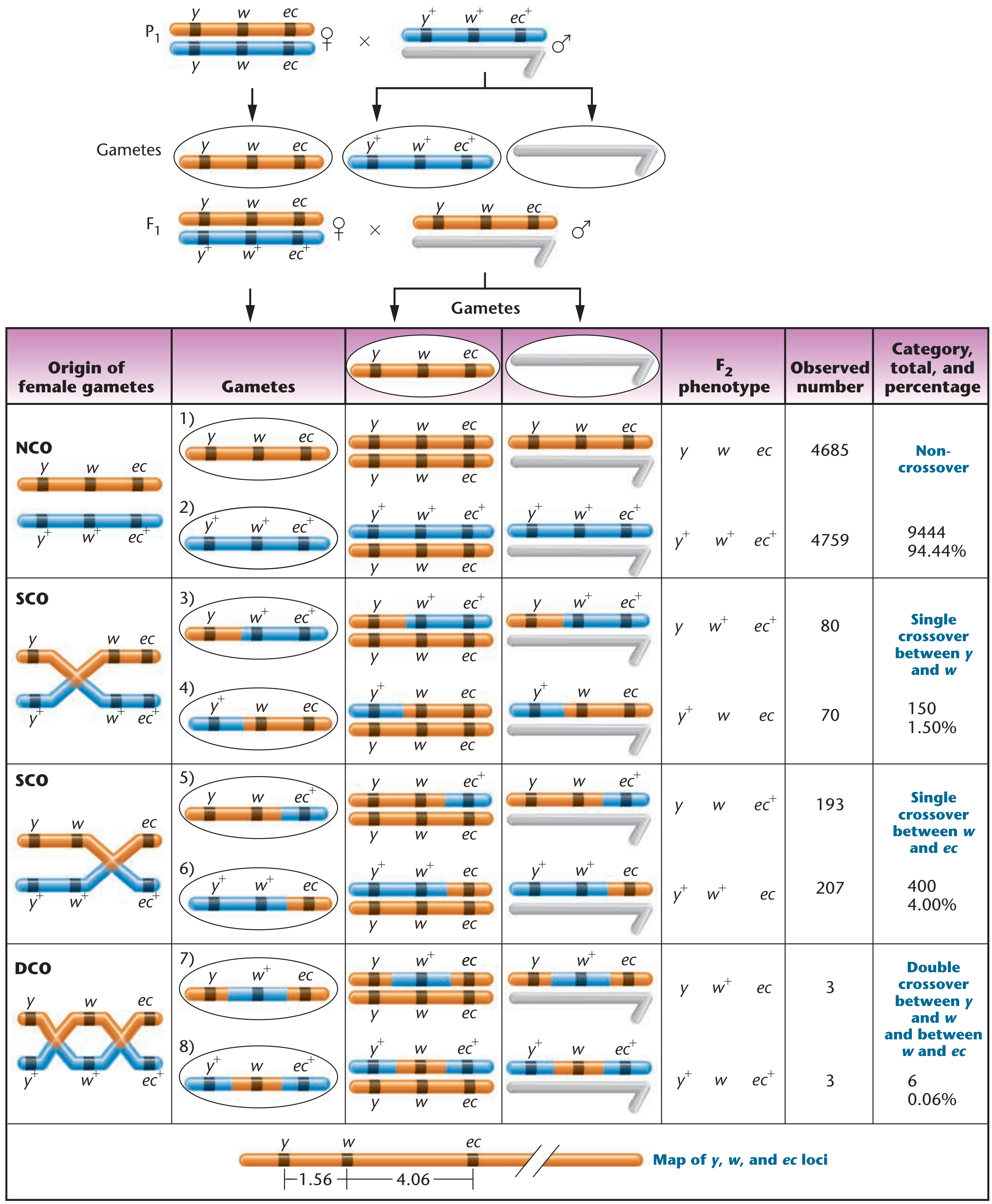
To execute a successful mapping cross, three criteria must be met:

1. The genotype of the organism producing the crossover gametes must be heterozygous at all loci under consideration. If homozygosity occurred at any locus, all gametes produced would contain the same allele, precluding mapping analysis.
2. The cross must be constructed so that the genotypes of all gametes can be accurately determined by observing the phenotypes of the resulting offspring. This is necessary because the gametes and their genotypes can never be observed directly. To overcome this problem, each phenotypic class must reflect the genotype of the gametes of the parents producing it.
3. A sufficient number of offspring must be produced in the mapping experiment to recover a representative sample of all crossover classes.

These criteria are met in the three-point mapping cross of *Drosophila melanogaster* shown in **Figure 5.8**. In this cross three X-linked recessive mutant genes—*yellow* body color, *white* eye color, and *echinus* eye shape—are considered. To diagram the cross, *we must assume some theoretical sequence, even though we do not yet know if it is correct.* In Figure 5.8, we initially assume the sequence of the three genes to be *y*—*w*—*ec*. If this is incorrect, our analysis shall demonstrate it and reveal the correct sequence.

In the P_1 generation, males hemizygous for all three wild-type alleles are crossed to females that are homozygous for all three recessive mutant alleles. Therefore, the P_1 males are wild type with respect to body color, eye color, and eye shape. They are said to have a *wild-type phenotype.* The females, on the other hand, exhibit the three mutant traits: yellow body color, white eyes, and echinus eye shape.

This cross produces an F_1 generation consisting of females that are heterozygous at all three loci and males that, because of the Y chromosome, are hemizygous for the three mutant alleles. Phenotypically, all F_1 females are wild type, while all F_1 males are yellow, white, and echinus. The genotype of the F_1 females fulfills the first criterion for constructing a map of the three linked genes; that is, it is heterozygous at the three loci and may serve as the source of recombinant gametes generated by crossing over. Note that, because of the genotypes of the P_1 parents, all three of the mutant alleles are on one homolog and all three wild-type alleles are on the other homolog. With other parents, *other arrangements would be possible that could produce a heterozygous genotype.* For example, a heterozygous female could have the *y* and *ec* mutant alleles on one homolog and the *w* allele on the other. This would occur if one of her parents was yellow, echinus and the other parent was white.



Origin of female gametes	Gametes	y w ec		F_2 phenotype	Observed number	Category, total, and percentage
NCO y w ec y^+ w^+ ec^+	1) y w ec	y w ec y w ec	y w ec	y w ec	4685	**Non-crossover**
	2) y^+ w^+ ec^+	y^+ w^+ ec^+ y w ec	y^+ w^+ ec^+	y^+ w^+ ec^+	4759	9444 94.44%
SCO y w ec y^+ w^+ ec^+	3) y w^+ ec^+	y w^+ ec^+ y w ec	y w^+ ec^+	y w^+ ec^+	80	**Single crossover between *y* and *w***
	4) y^+ w ec	y^+ w ec y w ec	y^+ w ec	y^+ w ec	70	150 1.50%
SCO y w ec y^+ w^+ ec^+	5) y w ec^+	y w ec^+ y w ec	y w ec^+	y w ec^+	193	**Single crossover between *w* and *ec***
	6) y^+ w^+ ec	y^+ w^+ ec y w ec	y^+ w^+ ec	y^+ w^+ ec	207	400 4.00%
DCO y w ec y^+ w^+ ec^+	7) y w^+ ec	y w^+ ec y w ec	y w^+ ec	y w^+ ec	3	**Double crossover between *y* and *w* and between *w* and *ec***
	8) y^+ w ec^+	y^+ w ec^+ y w ec	y^+ w ec^+	y^+ w ec^+	3	6 0.06%

FIGURE 5.8 A three-point mapping cross involving the *yellow* (y or y^+), *white* (w or w^+), and *echinus* (ec or ec^+) genes in *Drosophila melanogaster*. NCO, SCO, and DCO refer to noncrossover, single-crossover, and double-crossover groups, respectively. Centromeres are not drawn on the chromosomes, and only two nonsister chromatids are initially shown in the left-hand column.

In our cross, the second criterion is met as a result of the gametes formed by the F_1 males. Every gamete contains either an X chromosome bearing the three mutant alleles or a Y chromosome, which does not contain any of the three loci being considered. Whichever type participates in fertilization, the genotype of the gamete produced by the F_1 female will be expressed phenotypically in the F_2 female and male offspring derived from it. As a result, all noncrossover and crossover gametes produced by the F_1 female parent can be determined by observing the F_2 phenotypes.

With these two criteria met, we can construct a chromosome map from the crosses illustrated in Figure 5.8. First, we must determine which F_2 phenotypes correspond to the various noncrossover and crossover categories. To determine the **noncrossover** F_2 phenotypes, we must identify individuals derived from the parental gametes formed by the F_1 female. Each such gamete contains *an X chromosome unaffected by crossing over.* As a result of segregation, approximately equal proportions of the two types of gametes, and subsequently their F_2 phenotypes, are produced. Because they derive from a heterozygote, the genotypes of the two parental gametes and the F_2 phenotypes complement one another. For example, if one is wild type, the other is mutant for all three genes. This is the case in the cross being considered. In other situations, if one chromosome shows one mutant allele, the second chromosome shows the other two mutant alleles, and so on. These are therefore called **reciprocal classes** of gametes and phenotypes.

The two noncrossover phenotypes are most easily recognized because *they occur in the greatest proportion of offspring.* Figure 5.8 shows that gametes 1) and 2) are present in the greatest numbers. Therefore, flies that are yellow, white, and echinus and those that are normal, or wild type, for all three characters constitute the noncrossover category and represent 94.44 percent of the F_2 offspring.

The second category that can be easily detected is represented by the double-crossover phenotypes. Because of their low probability of occurrence, *they must be present in the least numbers.* Remember that this group represents two independent but simultaneous single-crossover events. Two reciprocal phenotypes can be identified: gamete 7), which shows the mutant traits yellow and echinus, but normal eye color; and gamete 8), which shows the mutant trait white, but normal body color and eye shape. Together these double-crossover phenotypes constitute only 0.06 percent of the F_2 offspring.

The remaining four phenotypic classes fall into two categories resulting from single crossovers. Gametes 3) and 4), reciprocal phenotypes produced by single-crossover events occurring between the *yellow* and *white* loci, are equal to 1.50 percent of the F_2 offspring. Gametes 5) and 6), constituting 4.00 percent of the F_2 offspring, represent the reciprocal phenotypes resulting from single-crossover events occurring between the *white* and *echinus* loci.

We can now calculate the map distances between the three loci. The distance between *y* and *w,* or between *w* and *ec,* is equal to the percentage of all detectable exchanges occurring between them. For any two genes under consideration, this includes all related single crossovers as well as all double crossovers. *The latter are included because they represent two simultaneous single crossovers.* For the *y* and *w* genes, this includes gametes 3), 4), 7), and 8), totaling 1.50% + 0.06%, or 1.56 mu. Similarly, the distance between *w* and *ec* is equal to the percentage of offspring resulting from an exchange between these two loci: gametes 5), 6), 7), and 8), totaling 4.00% + 0.06%, or 4.06 mu. The map of these three loci on the X chromosome is shown at the bottom of Figure 5.8.

Determining the Gene Sequence

In the preceding example, we assumed that the sequence (or order) of the three genes along the chromosome was *y—w—ec.* Our analysis established that the sequence is consistent with the data. However, in most mapping experiments, the gene sequence is not known, and this constitutes another variable in the analysis. In our example, had the gene order been unknown, we could have used one of two methods (which we will study next) to determine it. In your own work, you should select one of these methods and use it consistently.

Method I This method is based on the fact that there are only three possible arrangements, each containing a different one of the three genes between the other two:

(I)	*w—y—ec*	(*y* is in the middle)
(II)	*y—ec—w*	(*ec* is in the middle)
(III)	*y—w—ec*	(*w* is in the middle)

Use the following steps during your analysis to determine the gene order:

1. Assuming any of the three orders, first determine the *arrangement of alleles* along each homolog of the heterozygous parent giving rise to noncrossover and crossover gametes (the F_1 female in our example).
2. Determine whether a double-crossover event occurring within that arrangement will produce the *observed double-crossover phenotypes.* Remember that these phenotypes occur least frequently and are easily identified.
3. If this order does not produce the correct phenotypes, try each of the other two orders. One must work!

These steps are shown in **Figure 5.9**, using our *y—w—ec* cross. Three arrangements, labeled I, II, and III, are possible.

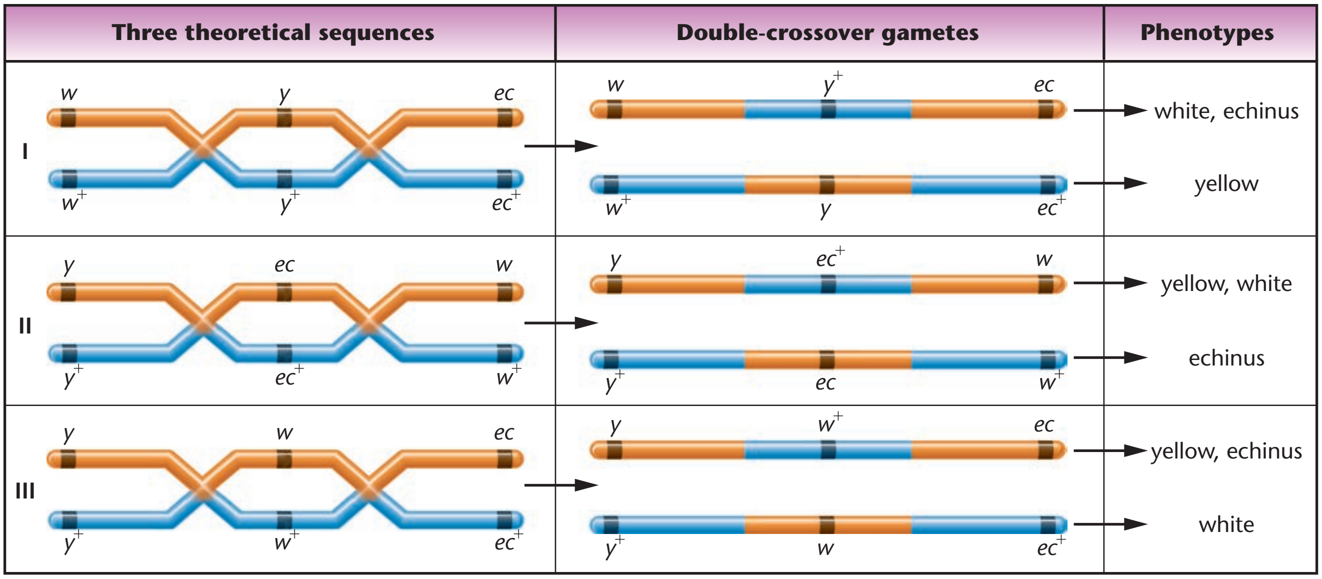


FIGURE 5.9 The three possible sequences of the *white, yellow,* and *echinus* genes, the results of a double crossover in each case, and the resulting phenotypes produced in a testcross. For simplicity, the two noncrossover chromatids of each tetrad are omitted.

1. Assuming that *y* is between *w* and *ec* (arrangement I), the distribution of alleles between the homologs of the F_1 heterozygote is

$$\frac{w \qquad y \qquad ec}{w^+ \qquad y^+ \qquad ec^+}$$

We know this because of the way in which the P_1 generation was crossed: The P_1 female contributes an X chromosome bearing the *w, y,* and *ec* alleles, while the P_1 male contributed an X chromosome bearing the w^+, y^+, and ec^+ alleles.

2. A double crossover within that arrangement yields the following gametes:

$$\underline{w \quad y^+ \quad ec} \quad \text{and} \quad \underline{w^+ \quad y \quad ec^+}$$

Following fertilization, if *y* is in the middle, the F_2 double-crossover phenotypes will correspond to these gametic genotypes, yielding offspring that express the white, echinus phenotype and offspring that express the yellow phenotype. Instead, determination of the actual double crossovers reveals them to be yellow, echinus flies and white flies. *Therefore, our assumed order is incorrect.*

3. If we consider arrangement II, with the ec/ec^+ alleles in the middle, or arrangement III, with the w/w^+ alleles in the middle:

$$\text{(II)} \ \frac{y \qquad ec \qquad w}{y^+ \qquad ec^+ \qquad w^+} \quad \text{or (III)} \ \frac{y \qquad w \qquad ec}{y^+ \qquad w^+ \qquad ec^+}$$

we see that arrangement II again provides predicted double-crossover phenotypes that do not correspond to the actual (observed) double-crossover phenotypes. The predicted phenotypes are yellow, white flies and echinus flies in the F_2 generation. *Therefore, this order is also incorrect.* However, arrangement III produces the observed phenotypes—yellow, echinus flies and white flies. *Therefore, this arrangement, with the w gene in the middle, is correct.*

To summarize Method I: First, determine the arrangement of alleles on the homologs of the heterozygote yielding the crossover gametes by identifying the reciprocal noncrossover phenotypes. Then, test each of the three possible orders to determine which one yields the observed double-crossover phenotypes—*the one that does so represents the correct order.* This method is summarized in Figure 5.9.

Method II Method II also begins by determining the arrangement of alleles along each homolog of the heterozygous parent. In addition, it requires one further assumption: *Following a double-crossover event, the allele in the middle position will fall between the outside, or flanking, alleles that were present on the opposite parental homolog.*

To illustrate, assume order I, *w—y—ec,* in the following arrangement:

$$\frac{w \qquad y \qquad ec}{w^+ \qquad y^+ \qquad ec^+}$$

Following a double-crossover event, the *y* and y^+ alleles would be switched to this arrangement:

$$\frac{w \qquad y^+ \qquad ec}{w^+ \qquad y \qquad ec^+}$$

After segregation, two gametes would be formed:

$$\underline{w \quad y^{+} \quad ec} \quad \text{and} \quad \underline{w^{+} \quad y \quad ec^{+}}$$

Because the genotype of the gamete will be expressed directly in the phenotype following fertilization, the double-crossover phenotypes will be:

white, echinus flies and yellow flies

Note that the *yellow* allele, assumed to be in the middle, is now associated with the two outside markers of the other homolog, w^{+} and ec^{+}. However, these predicted phenotypes do not coincide with the observed double-crossover phenotypes. Therefore, the *yellow* gene is not in the middle.

This same reasoning can be applied to the assumption that the *echinus* gene or the *white* gene is in the middle. In the former case, we will reach a negative conclusion. If we assume that the *white* gene is in the middle, the *predicted* and *actual* double crossovers coincide. Therefore, we conclude that the *white* gene is located between the *yellow* and *echinus* genes.

To summarize Method II, determine the arrangement of alleles on the homologs of the heterozygote yielding crossover gametes. Then examine the actual double-crossover phenotypes and identify the single allele that has been switched so that it is now no longer associated with its original neighboring alleles. That allele will be the one located between the other two in the sequence.

In our example *y, ec,* and *w* are on one homolog in the F_1 heterozygote, and y^{+}, ec^{+}, and w^{+} are on the other. In the F_2 double-crossover classes, it is *w* and w^{+} that have been switched. The *w* allele is now associated with y^{+} and ec^{+}, while the w^{+} allele is now associated with the *y* and *ec* alleles. Therefore, the *white* gene is in the middle, and the *yellow* and *echinus* genes are the flanking markers.

An Autosomal Mapping Problem in Maize

Having established the basic principles of chromosome mapping, we will now consider a related problem in maize (corn). This analysis differs from the preceding example in two ways. First, the previous mapping cross involved X-linked genes. Here, we consider autosomal genes. Second, in the discussion of this cross, we will change our use of symbols (as first suggested in Chapter 4). Instead of using the gene symbols and superscripts (e.g., bm^{+}, v^{+}, and pr^{+}), we simply use + to denote each wild-type allele. This system is easier to manipulate but requires a better understanding of mapping procedures.

When we look at three autosomally linked genes in maize, our experimental cross must still meet the same three criteria we established for the X-linked genes in *Drosophila:* (1) One parent must be heterozygous for all traits under consideration; (2) the gametic genotypes produced by the heterozygote must be apparent from observing the phenotypes of the offspring; and (3) a sufficient sample size must be available for complete analysis.

In maize, the recessive mutant genes *bm* (*brown* midrib), *v* (*virescent* seedling), and *pr* (*purple* aleurone) are linked on chromosome 5. Assume that a female plant is known to be heterozygous for all three traits, but we do not know: (1) the arrangement of the mutant alleles on the maternal and paternal homologs of this heterozygote; (2) the sequence of genes; or (3) the map distances between the genes. What genotype must the male plant have to allow successful mapping? To meet the second criterion, the male must be homozygous for all three recessive mutant alleles. Otherwise, offspring of this cross showing a given phenotype might represent more than one genotype, making accurate mapping impossible. Note that this is equivalent to performing a testcross.

Figure 5.10 diagrams this cross. As shown, we know neither the *arrangement of alleles* nor the *sequence of loci* in the heterozygous female. Several possibilities are shown, but we have yet to determine which is correct. We don't know the sequence in the testcross male parent either, so we must designate it randomly. Note that we initially placed *v* in the middle. *This may or may not be correct.*

The offspring have been arranged in groups of two, representing each pair of reciprocal phenotypic classes. The four reciprocal classes are derived from no crossing over (NCO), each of two possible single-crossover events (SCO), and a double-crossover event (DCO).

To solve this problem, refer to Figures 5.10 and 5.11 as you consider the following questions:

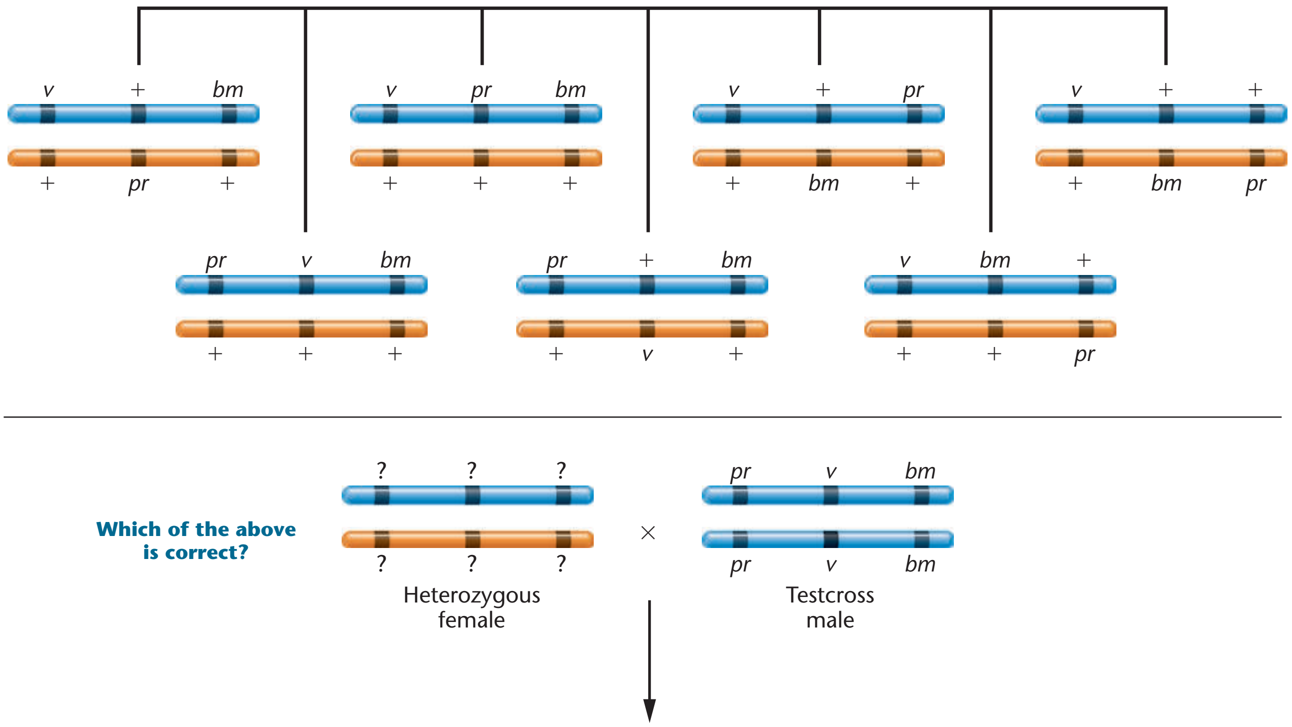
1. *What is the correct heterozygous arrangement of alleles in the female parent?*
   Determine the two noncrossover classes, those that occur with the highest frequency. In this case, they are $\underline{+\ v\ bm}$ and $\underline{pr\ +\ +}$. Therefore, the alleles on the homologs of the female parent must be distributed as shown in **Figure 5.11(a)**. These homologs segregate into gametes, unaffected by any recombination event. Any other arrangement of alleles will not yield the observed noncrossover classes. (Remember that $\underline{+\ v\ bm}$ is equivalent to $\underline{pr^{+}\ v\ bm}$ and that $\underline{pr\ +\ +}$ is equivalent to $\underline{pr\ v^{+}\ bm^{+}}$.)

2. *What is the correct sequence of genes?*
   To answer this question, we will first use the approach described in Method I. We know, based on the answer to question 1, that the correct arrangement of alleles is

$$\frac{\underline{+ \qquad v \qquad bm}}{pr \qquad + \qquad +}$$

But is the gene sequence correct? That is, will a double-crossover event yield the observed double-crossover

(a) Some possible allele arrangements and gene sequences in a heterozygous female



(b) Actual results of mapping cross*

Phenotypes of offspring	Number	Total and percentage	Exchange classification
+ v bm pr + +	230 237	467 42.1%	Noncrossover (NCO)
+ + bm pr v +	82 79	161 14.5%	Single crossover (SCO)
+ v + pr + bm	200 195	395 35.6%	Single crossover (SCO)
pr v bm + + +	44 42	86 7.8%	Double crossover (DCO)

* The sequence *pr* – *v* – *bm* may or may not be correct.

FIGURE 5.10 (a) Some possible allele arrangements and gene sequences in a heterozygous female. The data from a three-point mapping cross, depicted in (b), where the female is testcrossed, provide the basis for determining which combination of arrangement and sequence is correct. [See Figure 5.11(d).]

phenotypes following fertilization? *Observation shows that it will not* [Figure 5.11(b)]. Now try the other two orders [Figure 5.11(c) and (d)], *keeping the same allelic arrangement*:

$$\frac{+ \qquad bm \qquad v}{pr \qquad + \qquad +} \quad \text{or} \quad \frac{v \qquad + \qquad bm}{+ \qquad pr \qquad +}$$

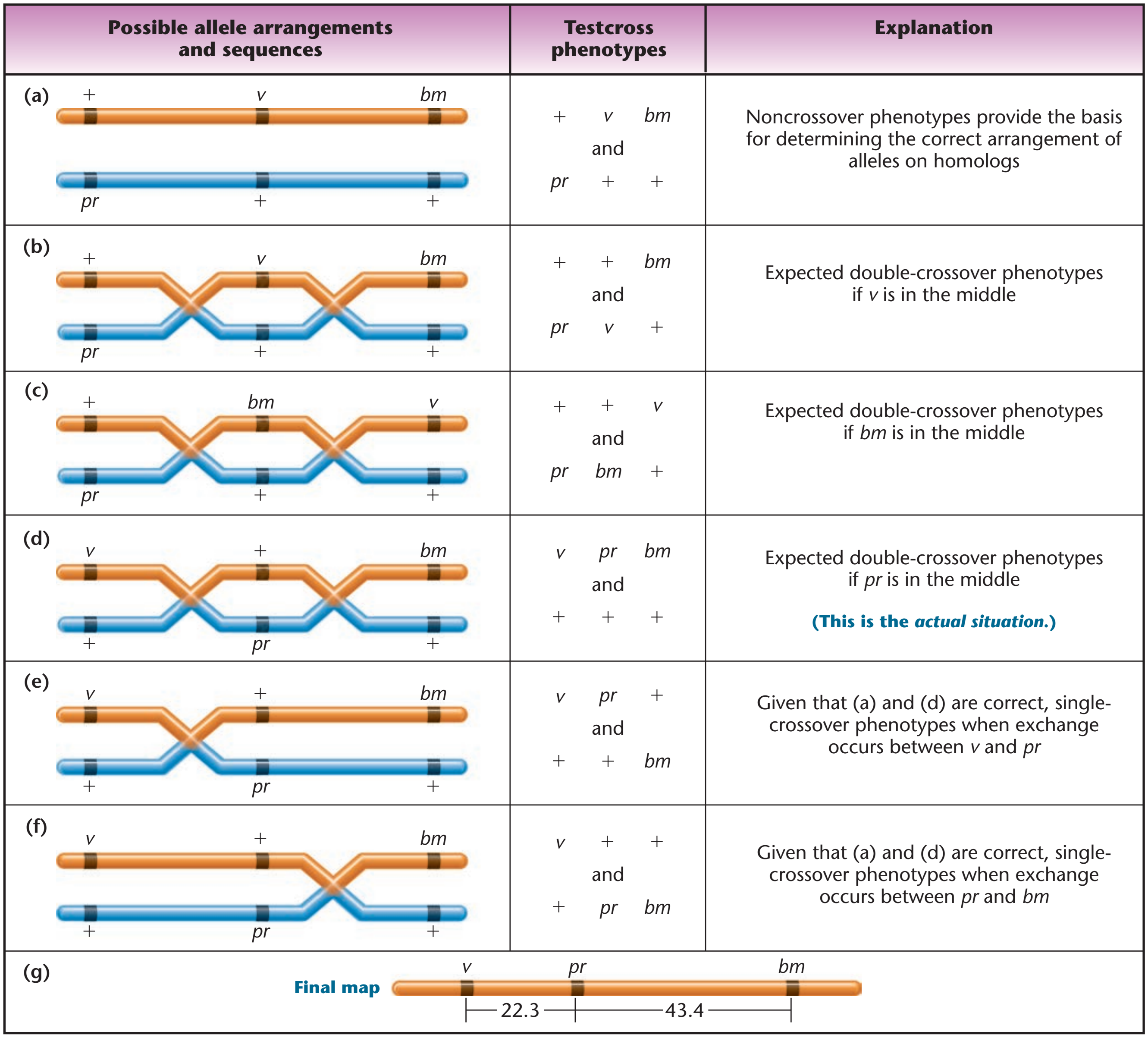
Only the order on the right yields the observed double-crossover gametes [Figure 5.11(d)]. Therefore, the *pr* gene is in the middle.

The same conclusion is reached if we use Method II to analyze the problem. In this case, no assumption of gene sequence is necessary. The arrangement of alleles along homologs in the heterozygous parent is

$$\frac{+ \qquad v \qquad bm}{pr \qquad + \qquad +}$$

The double-crossover gametes are also known:

$$\underline{pr \quad v \quad bm} \quad \text{and} \quad \underline{+ \quad + \quad +}$$



Possible allele arrangements and sequences	Testcross phenotypes	Explanation
(a)	+ v bm and pr + +	Noncrossover phenotypes provide the basis for determining the correct arrangement of alleles on homologs
(b)	+ + bm and pr v +	Expected double-crossover phenotypes if *v* is in the middle
(c)	+ + v and pr bm +	Expected double-crossover phenotypes if *bm* is in the middle
(d)	v pr bm and + + +	Expected double-crossover phenotypes if *pr* is in the middle **(This is the *actual situation*.)**
(e)	v pr + and + + bm	Given that (a) and (d) are correct, single-crossover phenotypes when exchange occurs between *v* and *pr*
(f)	v + + and + pr bm	Given that (a) and (d) are correct, single-crossover phenotypes when exchange occurs between *pr* and *bm*
(g)		

FIGURE 5.11 Producing a map of the three genes in the cross in Figure 5.10, where neither the arrangement of alleles nor the sequence of genes in the heterozygous female parent is known.

We can see that it is the *pr* allele that has shifted relative to its noncrossover arrangement, so as to be associated with *v* and *bm* following a double crossover. The latter two alleles (*v* and *bm*) were present together on one homolog, and they stayed together. Therefore, *pr* is the odd gene, so to speak, and is located in the middle. Thus, we arrive at the same arrangement and sequence as we did with Method I:

$$\frac{v \quad + \quad bm}{+ \quad pr \quad +}$$

3. *What is the distance between each pair of genes?*
Having established the correct sequence of loci as *v—pr—bm,* we can now determine the distance between *v* and *pr* and between *pr* and *bm*. Remember that the map distance between two genes is calculated on the basis of all detectable recombinational events occurring between them. This includes both the single and double-crossover events.

Figure 5.11(e) shows that the phenotypes $\underline{v\ pr\ +}$ and $\underline{+ + bm}$ result from single crossovers between *v* and *pr,* and Figure 5.10 shows that those single crossovers account for 14.5 percent of the offspring. By adding the percentage of double crossovers (7.8 percent) to the number obtained for those single crossovers, we calculate the total distance between *v* and *pr* to be 22.3 mu.

Figure 5.11(f) shows that the phenotypes $\underline{v + +}$ and $\underline{+ \, pr \, bm}$ result from single crossovers between the *pr* and *bm* loci, totaling 35.6 percent, according to Figure 5.10. Adding the double-crossover classes (7.8 percent), we compute the distance between *pr* and *bm* as 43.4 mu. The final map for all three genes in this example is shown in **Figure 5.11(g)**.

NOW SOLVE THIS

5.3 In *Drosophila*, a heterozygous female for the X-linked recessive traits *a*, *b*, and *c* was crossed to a male that phenotypically expressed *a*, *b*, and *c*. The offspring occurred in the following phenotypic ratios.

+	*b*	*c*	460
a	+	+	450
a	*b*	*c*	32
+	+	+	38
a	+	*c*	11
+	*b*	+	9

No other phenotypes were observed.

(a) What is the genotypic arrangement of the alleles of these genes on the X chromosome of the female?

(b) Determine the correct sequence and construct a map of these genes on the X chromosome.

(c) What progeny phenotypes are missing? Why?

■ **HINT:** *This problem involves a three-point mapping experiment where only six phenotypic categories are observed, even though eight categories are typical of such a cross. The key to its solution is to be aware that if the distances between the loci are relatively small, the sample size may be too small for the predicted number of double crossovers to be recovered, even though reciprocal pairs of single crossovers are seen. You should write the missing gametes down as double crossovers and record zeros for their frequency of appearance.*

For more practice, see Problems 15–17.

5.4 As the Distance between Two Genes Increases, Mapping Estimates Become More Inaccurate

So far, we have assumed that crossover frequencies are directly proportional to the distance between any two loci along the chromosome. However, it is not always possible to detect all crossover events. A case in point is a double exchange that occurs between the two loci in question. As shown in **Figure 5.12(a)**, if a double exchange occurs, the original arrangement of alleles on each nonsister homolog is recovered. Therefore, even though crossing over has occurred, it is impossible to detect. This phenomenon is true for all even-numbered exchanges between two loci.

Furthermore, as a result of complications posed by multiple-strand exchanges, mapping determinations usually underestimate the actual distance between two genes. The farther apart two genes are, the greater the probability that undetected crossovers will occur. While the discrepancy is minimal for two genes relatively close together, the degree of inaccuracy increases as the distance increases, as shown in the graph of recombination frequency versus map distance in **Figure 5.12(b)**. There, the theoretical frequency where a direct correlation between recombination and map distance exists is contrasted with the actual frequency observed as the distance between two genes increases. The most accurate maps are constructed from experiments in which genes are relatively close together.

Interference and the Coefficient of Coincidence

As review of the product law in Section 5.3 would indicate, the expected frequency of multiple exchanges, such as double crossovers, can be predicted once the distance between genes is established. For example, in the maize cross of Section 5.3, the distance between *v* and *pr* is 22.3 mu, and the distance between *pr* and *bm* is 43.4 mu. If the two single crossovers that make up a double crossover occur independently of one another, we can calculate the expected frequency of double crossovers ($\mathrm{DCO_{exp}}$) as follows:

$$\mathrm{DCO_{exp}} = (0.223) \times (0.434) = 0.097 = 9.7\%$$

Often in mapping experiments, the observed DCO frequency is less than the expected number of DCOs. In the maize cross, for example, only 7.8 percent of the DCOs are observed when 9.7 percent are expected. **Interference (*I*),** the inhibition of further crossover events by a crossover event in a nearby region of the chromosome, causes this reduction.

To quantify the disparities that result from interference, we calculate the **coefficient of coincidence (*C*)**:

$$C = \frac{\text{Observed DCO}}{\text{Expected DCO}}$$

In the maize cross, we have

$$C = \frac{0.078}{0.097} = 0.804$$

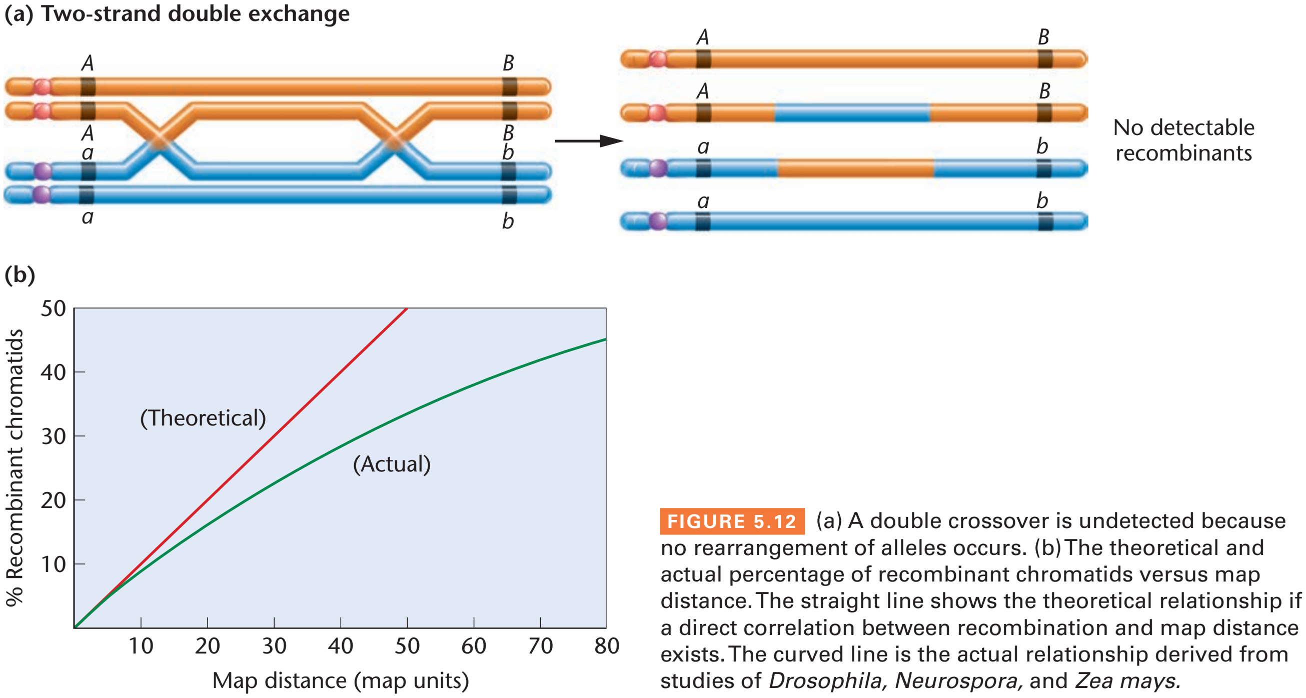


FIGURE 5.12 (a) A double crossover is undetected because no rearrangement of alleles occurs. (b) The theoretical and actual percentage of recombinant chromatids versus map distance. The straight line shows the theoretical relationship if a direct correlation between recombination and map distance exists. The curved line is the actual relationship derived from studies of *Drosophila, Neurospora,* and *Zea mays.*

Once we have found C, we can quantify interference (I) by using this simple equation:

$$I = 1 - C$$

In the maize cross, we have

$$I = 1.000 - 0.804 = 0.196$$

If interference is complete and no double crossovers occur, then $I = 1.0$. If fewer DCOs than expected occur, I is a positive number and **positive interference** has occurred. If more DCOs than expected occur, I is a negative number and **negative interference** has occurred. In this example, I is a positive number (0.196), indicating that 19.6 percent fewer double crossovers occurred than expected.

Positive interference is most often observed in eukaryotic systems. In *C. elegans,* for example, only one crossover event per chromosome is observed, and interference along each chromosome is complete ($C = 1.0$). In other organisms, the closer genes are to one another along the chromosome, the more positive interference occurs. Interference in *Drosophila* is often complete within a distance of 10 map units. This observation suggests that physical constraints preventing the formation of closely spaced chiasmata contribute to interference. The interpretation is consistent with the finding that interference decreases as the genes in question are located farther apart. In the maize cross illustrated in Figures 5.10 and 5.11, the three genes are relatively far apart, and 80 percent of the expected double crossovers are observed.

5.5 *Drosophila* Genes Have Been Extensively Mapped

In organisms such as fruit flies, maize, and the mouse, where large numbers of mutants have been discovered and where mapping crosses are possible, extensive maps of each chromosome have been constructed. **Figure 5.13** presents partial maps of the four chromosomes of *Drosophila melanogaster.* Virtually every morphological feature of the fruit fly has been subjected to mutation. Each locus affected by mutation is first localized to one of the four chromosomes, or linkage groups, and then mapped in relation to other genes present on that chromosome. As you can see, the genetic map of the X chromosome is somewhat shorter than that of autosome II or III. In comparison to these three, autosome IV is miniscule. Cytological evidence has shown that the relative lengths of the genetic maps correlate roughly with the relative physical lengths of these chromosomes.

EVOLVING CONCEPT OF THE GENE

Based on the gene-mapping studies in *Drosophila* and many other organisms from the 1920s through the mid-1950s, geneticists regarded genes as hereditary units organized in a specific sequence along chromosomes, between which recombination could occur. Genes were thus viewed as indivisible “beads on a string.” ■

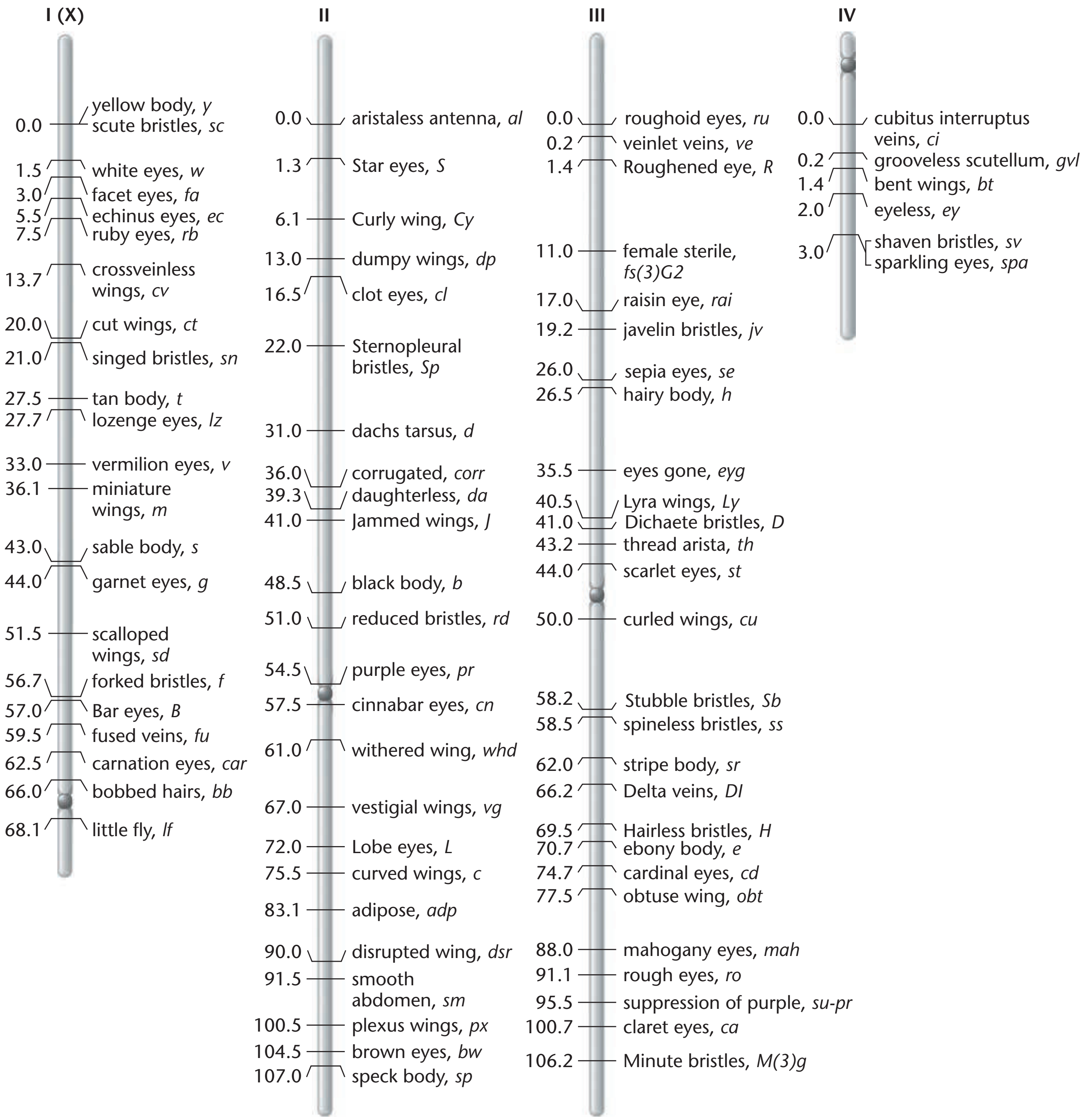


FIGURE 5.13 A partial genetic map of the four chromosomes of *Drosophila melanogaster*. The circle on each chromosome represents the position of the centromere. Chromosome I is the X chromosome. Chromosome IV is not drawn to scale; that is, it is relatively smaller than indicated.

5.6 Lod Score Analysis and Somatic Cell Hybridization Were Historically Important in Creating Human Chromosome Maps

In humans, genetic experiments involving carefully planned crosses and large numbers of offspring are neither ethical nor feasible, so the earliest linkage studies were based on pedigree analysis. These studies attempted to establish whether certain traits were X-linked or autosomal. As we showed earlier in the text (see Chapter 4), traits determined by genes located on the X chromosome result in characteristic pedigrees; thus, such genes were easier to identify. For autosomal traits, geneticists tried to distinguish clearly whether pairs of traits demonstrated linkage or independent assortment. When extensive pedigrees are available, it is possible to conclude that two genes under consideration are closely linked (i.e., rarely separated by crossing over) from the fact that the two traits segregate together. This approach established linkage between the genes encoding the **Rh antigens** and the gene responsible for the phenotype referred to as **elliptocytosis,** where the shape of erythrocytes is oval. It was hoped that from these kinds of observations a human gene map could be created.

A difficulty arises, however, when two genes of interest are separated on a chromosome to the degree that recombinant gametes are formed, obscuring linkage in a pedigree. In these cases, an approach relying on probability calculations, called the **lod score method,** helps to demonstrate linkage. First devised by J. B. S. Haldane and C. A. Smith in 1947 and refined by Newton Morton in 1955, the lod score (standing for *l*og of the *od*ds favoring linkage) assesses the probability that a particular pedigree (or several pedigrees for the same traits of interest) involving two traits reflects genetic linkage between them. First, the probability is calculated that the family (pedigree) data concerning two traits conform to transmission without linkage—that is, the traits appear to be independently assorting. Then the probability is calculated that the identical family data for these same traits result from linkage with a specified recombination frequency. These probability calculations factor in the statistical significance at the $p = 0.05$ level. The ratio of these probability values is then calculated and converted to the logarithm of this value, which reflects the "odds" for, and against, linkage. Traditionally, a value of 3.0 or higher strongly indicates linkage, whereas a value of −2.0 or less argues strongly against linkage. Values between −2.0 and 3.0 are considered to be inconclusive.

The lod score method represented an important advance in assigning human genes to specific chromosomes and in constructing preliminary human chromosome maps. However, its accuracy was limited by the extent of the pedigree, and the initial results were discouraging—both because of this limitation and because of the relatively high haploid number of human chromosomes (23). By 1960, very little autosomal linkage information had become available.

However, in the 1960s, a new technique, **somatic cell hybridization,** proved to be an immense aid in assigning human genes to their respective chromosomes. This technique, first discovered by Georges Barski, relies on the fact that two cells in culture can be induced to fuse into a single hybrid cell. Barski used two mouse-cell lines, but it soon became evident that cells from different organisms could also be fused. When fusion occurs, an initial cell type called a **heterokaryon** is produced. The hybrid cell contains two nuclei in a common cytoplasm. Using the proper techniques, we can fuse human and mouse cells, for example, and isolate the hybrids from the parental cells.

As the heterokaryons are cultured *in vitro,* two interesting changes occur. Eventually, the nuclei fuse together, creating a **synkaryon.** Then, as culturing is continued for many generations, chromosomes from one of the two parental species are gradually lost. In the case of the human–mouse hybrid, human chromosomes are lost randomly until eventually the synkaryon has a full complement of mouse chromosomes and only a few human chromosomes. It is the preferential loss of human chromosomes (rather than mouse chromosomes) that makes possible the assignment of human genes to the chromosomes on which they reside.

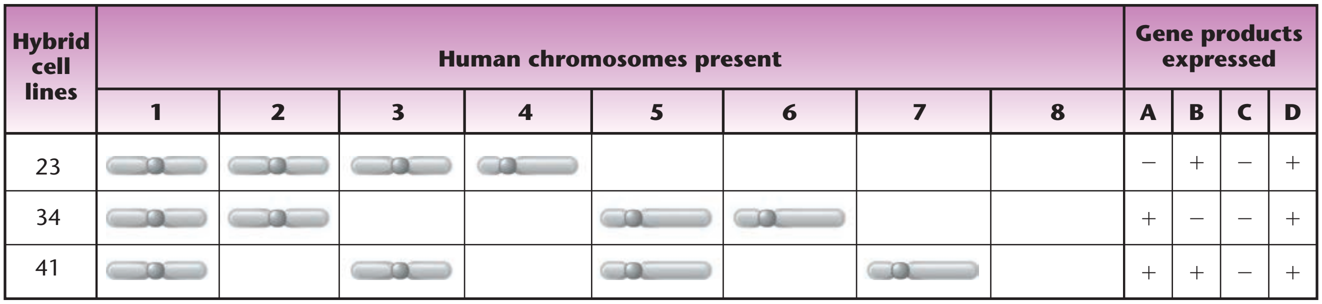
The experimental rationale is straightforward. If a specific human gene product is synthesized in a synkaryon containing three human chromosomes, for example, then the gene responsible for that product must reside on one of the three human chromosomes remaining in the hybrid cell. On the other hand, if the human gene product is not synthesized in the synkaryon, the responsible gene cannot be present on any of the remaining three human chromosomes. Ideally, one would have a panel of 23 hybrid cell lines, each with a different human chromosome, allowing the immediate assignment to a particular chromosome of any human gene for which the product could be characterized.

In practice, a panel of cell lines each of which contains several remaining human chromosomes is most often used. The correlation of the presence or absence of each chromosome with the presence or absence of each gene product is called **synteny testing.** Consider, for example, the hypothetical data provided in **Figure 5.14**, where four gene products (A, B, C, and D) are tested in relationship to eight human chromosomes. Let us carefully analyze the results to locate the gene that produces product A.

1. Product A is not produced by cell line 23, but chromosomes 1, 2, 3, and 4 are present in cell line 23. Therefore, we can rule out the presence of gene *A* on those four chromosomes and conclude that it might be on chromosome 5, 6, 7, or 8.
2. Product A is produced by cell line 34, which contains chromosomes 5 and 6, but not 7 and 8. Therefore, gene *A* is on chromosome 5 or 6, but cannot be on 7 or 8 because they are absent, even though product A is produced.
3. Product A is also produced by cell line 41, which contains chromosome 5 but not chromosome 6. Therefore, gene *A* is on chromosome 5, according to this analysis.

Using a similar approach, we can assign gene *B* to chromosome 3. Perform the analysis for yourself to demonstrate that this is correct.

Gene *C* presents a unique situation. The data indicate that it is not present on chromosomes 1–7. While it might be on chromosome 8, no direct evidence supports this conclusion. Other panels are needed. We leave gene *D* for you to analyze. Upon what chromosome does it reside?



Hybrid cell lines	Human chromosomes present								Gene products expressed			
	1	2	3	4	5	6	7	8	A	B	C	D
23	▬	▬	▬	▬					−	+	−	+
34	▬	▬			▬	▬			+	−	−	+
41	▬		▬		▬		▬		+	+	−	+

FIGURE 5.14 A hypothetical grid of data used in synteny testing to assign genes to their appropriate human chromosomes. Three somatic hybrid cell lines, designated 23, 34, and 41, have each been scored for the presence, or absence, of human chromosomes 1 through 8, as well as for their ability to produce the hypothetical human gene products A, B, C, and D.

By using the approach just described, researchers were able to assign literally hundreds of human genes to one chromosome or another. To map genes for which the products have yet to be discovered, researchers have had to rely on other approaches. For example, by combining recombinant DNA technology with pedigree analysis, it was possible to assign the genes responsible for Huntington disease, cystic fibrosis, and neurofibromatosis to their respective chromosomes 4, 7, and 17. Modern genomic analysis has expanded our knowledge of the mapping location of countless other human traits, as described in Section 5.7.

5.7 Chromosome Mapping Is Currently Performed Using DNA Markers and Annotated Computer Databases

Although traditional methods based on recombination analysis produced detailed chromosomal maps in several organisms, maps in other organisms (including humans) that did not lend themselves to such studies were greatly limited. But that all changed rapidly with the development of recombinant DNA techniques, the discovery of molecular DNA markers, and most recently, our ability to sequence DNA and perform genomic analysis. These advances have greatly enhanced mapping in those organisms. We will address this topic using humans as an example.

Progress has initially relied on the discovery of **DNA markers** (mentioned earlier) that have been identified during recombinant DNA and genomic studies. These markers are short segments of DNA whose sequence and location are known, making them useful landmarks for mapping purposes. The analysis of human genes in relation to these markers has extended our knowledge of the location within the genome of countless genes, which is the ultimate goal of mapping.

The earliest examples are the DNA markers referred to as **restriction fragment length polymorphisms (RFLPs)** (see Chapter 22) and **microsatellites** or **short tandem repeats** (see Chapter 12). RFLPs are polymorphic sites generated when specific DNA sequences are recognized and cut by restriction enzymes. Microsatellites are short repetitive sequences that are found throughout the genome, and they vary in the number of repeats at any given site. For example, the two-nucleotide sequence CA is repeated 5–50 times per site $[(CA)_n]$ and appears throughout the genome approximately every 10,000 bases, on average. Microsatellites may be identified not only by the number of repeats but by the DNA sequences that flank them. More recently, variation in single nucleotides, called **single-nucleotide polymorphisms** (**SNPs,** also called "snips"), has been utilized. Found throughout the genome, up to several million of these variations may be screened for an association with a disease or trait of interest, thus providing geneticists with a means to identify and locate related genes.

Cystic fibrosis offers an early example of a gene located by using DNA markers. It is a life-shortening autosomal recessive exocrine disorder resulting in excessive, thick mucus that impedes the function of organs such as the lung and pancreas. After scientists established that the gene causing this disorder is located on chromosome 7, they were then able to pinpoint its exact location on the long arm (the q arm) of that chromosome.

In 2007, using SNPs as DNA markers, associations between 24 genomic locations were established with seven common human diseases: *Type 1* (insulin dependent) and *Type 2 diabetes, Crohn's disease* (inflammatory bowel disease), *hypertension, coronary artery disease, bipolar* (manic-depressive) *disorder,* and *rheumatoid arthritis.* In each case, an inherited susceptibility effect was mapped to a specific location on a specific chromosome within the genome. In some cases, this either confirmed or led to the identification of a specific gene involved in the cause of the disease.

During the past several decades, dramatic improvements in DNA sequencing technology have resulted in a proliferation of **sequence maps** for humans and many other species. Sequence maps provide the finest level of mapping detail because they pinpoint the nucleotide sequence of genes (and noncoding sequences) on a chromosome. The Human Genome Project resulted in sequence maps for all human chromosomes, providing an incredible level of detail about human gene sequences, the specific location of genes on a chromosome, and the proximity of genes and noncoding sequences to each other, among other details. For instance, when human chromosome sequences were analyzed by software programs, an approach called **bioinformatics**, to be discussed later in the text (see Chapter 22), geneticists could utilize such data to map possible protein-coding sequences in the genome. This led to the identification of thousands of potential genes that were previously unknown.

The many Human Genome Project databases that are now available (see the Exploring Genomics feature at the end of this chapter) make it possible to map genes along a human chromosome in base-pair distances rather than recombination frequency. This distinguishes what is referred to as a *physical map* of the genome from the *genetic maps* described earlier. When the genome sequence of a species is available, mapping by linkage or other genetic mapping approaches becomes obsolete.

5.8 Crossing Over Involves a Physical Exchange between Chromatids

Once genetic mapping techniques had been developed, they were used to study the relationship between the chiasmata observed in meiotic prophase I and crossing over. For example, are chiasmata visible manifestations of crossover events? If so, then crossing over in higher organisms appears to be the result of an actual physical exchange between homologous chromosomes. That this is the case was demonstrated independently in the 1930s by Harriet Creighton and Barbara McClintock in *Zea mays* (maize) and by Curt Stern in *Drosophila*.

Because the experiments are similar, we will consider only one of them, the work with maize. Creighton and McClintock studied two linked genes on chromosome 9 of the maize plant. At one locus, the alleles *colorless* (*c*) and *colored* (*C*) control endosperm coloration (the endosperm is the nutritive tissue inside the corn kernel). At the other locus, the alleles *starchy* (*Wx*) and *waxy* (*wx*) control the carbohydrate characteristics of the endosperm. The maize plant studied was heterozygous at both loci. The key to this experiment is that one of the homologs contained two unique cytological markers. The markers consisted of a densely stained knob at one end of the chromosome and a translocated piece of another chromosome (8) at the other end. The arrangements of these alleles and markers could be detected cytologically and are shown in **Figure 5.15**.

Creighton and McClintock crossed this plant to one homozygous for the *colorless* allele (*c*) and heterozygous for the *waxy/starchy* alleles. They obtained a variety of different phenotypes in the offspring, but they were most interested in one that occurred as a result of a crossover involving the chromosome with the unique cytological markers. They examined the chromosomes of this plant, having a colorless, waxy phenotype (Case I in Figure 5.15), for the presence of the cytological markers. If genetic crossing over was accompanied by a physical exchange between homologs, the translocated chromosome would still be present, but the knob would not. This was the case!

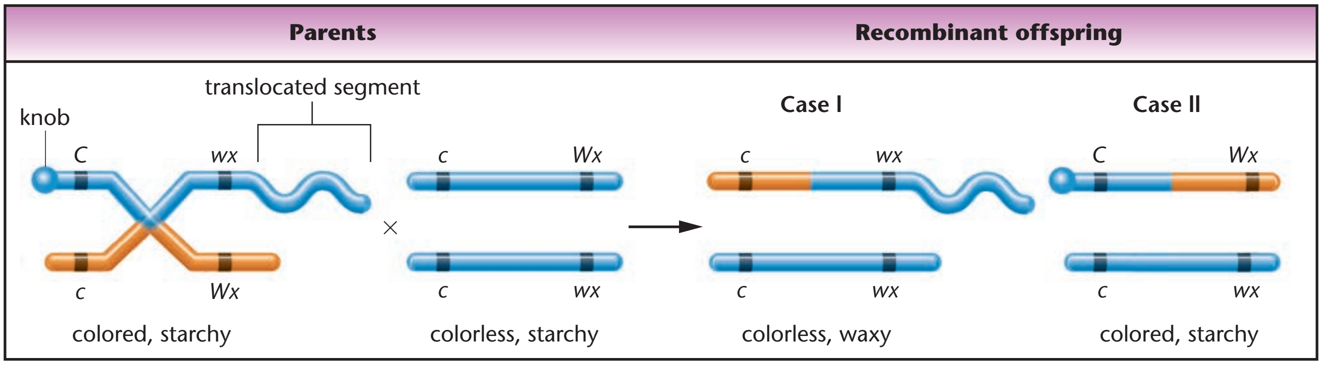


FIGURE 5.15 The phenotypes and chromosome compositions of parents and recombinant offspring in Creighton and McClintock's experiment in maize. The knob and translocated segment served as cytological markers, which established that crossing over involves an actual exchange of chromosome arms.

In a second plant (Case II), the phenotype colored, starchy should result from either nonrecombinant gametes or crossing over. Some of the cases then ought to contain chromosomes with the dense knob but not the translocated chromosome. This condition was also found, and the conclusion that a physical exchange had taken place was again supported. Along with Curt Stern's findings in *Drosophila,* this work clearly established that crossing over has a cytological basis.

Once we have discussed DNA in more detail (see Chapter 10), we will return to the topic of crossing over to examine how breakage and reunion occur between strands of DNA in chromatids. This discussion (see Chapter 11) will provide a better understanding of genetic recombination.

5.9 Exchanges Also Occur between Sister Chromatids during Mitosis

Considering that crossing over occurs between synapsed homologs in meiosis, we might ask whether such a physical exchange occurs between sister chromatids that are aligned together during mitosis. Each individual chromosome in prophase and metaphase of mitosis consists of two identical sister chromatids, joined at a common centromere. A number of experiments have demonstrated that reciprocal exchanges similar to crossing over do occur between sister chromatids. While these **sister chromatid exchanges (SCEs)** do not produce new allelic combinations, evidence is accumulating that attaches significance to these events.

Identification and study of SCEs are facilitated by several unique staining techniques. In each approach, cells are allowed to replicate for two generations in the presence of a base analog such as bromodeoxyuridine (BrdU, an analog of thymidine). Following two rounds of replication, each pair of sister chromatids has one member with one strand of DNA "labeled" with the analog and the other member with both strands labeled with it. Using a differential stain, chromatids with the analog in both strands stain *less brightly* than chromatids with BrdU in only one strand. As a result, any SCEs are readily detectable. In **Figure 5.16**, numerous instances of SCE events are clearly evident. Because of their patterns of alternating patches, these sister chromatids are sometimes referred to as **harlequin chromosomes.**

The significance of SCEs is still uncertain, but several observations have led to great interest in this phenomenon. We know, for example, that agents that induce chromosome damage (e.g., viruses, X rays, ultraviolet light, and certain chemical mutagens) also increase the frequency of SCEs. Further, the frequency of SCEs is elevated in **Bloom syndrome,** (see Figure 5.16), a human disorder caused by a mutation in the *BLM* gene on chromosome 15. This rare, recessively inherited disease is characterized by prenatal and postnatal retardation of growth, a great sensitivity of the facial skin to the sun, immune deficiency, a predisposition to malignant and benign tumors, and abnormal behavior patterns. The chromosomes from cultured leukocytes, bone marrow cells, and fibroblasts derived from homozygotes are very fragile and unstable when compared with those derived from homozygous and heterozygous normal individuals. Increased breaks and rearrangements between nonhomologous chromosomes are observed in addition to excessive amounts of SCEs. Work by James German and colleagues suggests that the *BLM* gene encodes an enzyme called **DNA helicase,** which is best known for its role in DNA replication (see Chapter 11).

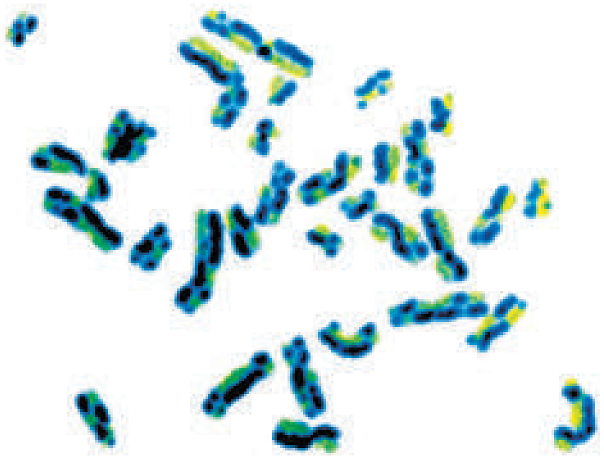
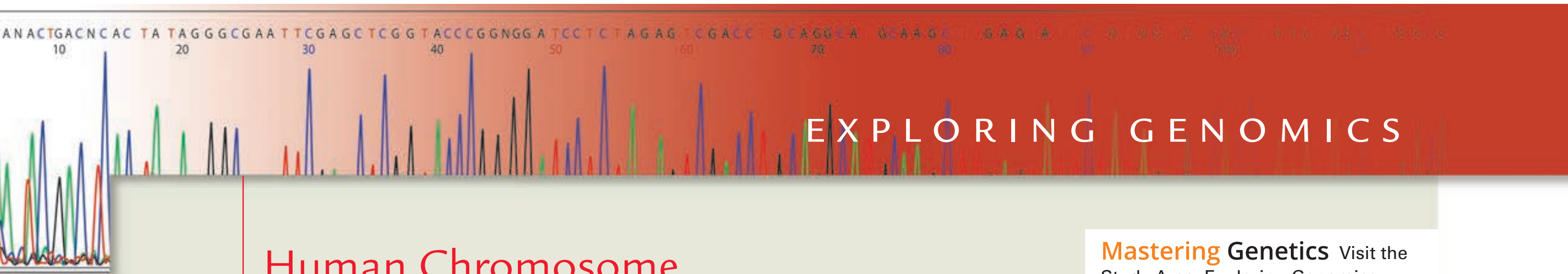


FIGURE 5.16 Demonstration of sister chromatid exchanges (SCEs) in mitotic chromosomes from a Bloom syndrome patient, which display elevated numbers of SCEs. Chromosomes with SCEs are sometimes called harlequin chromosomes because of the alternating patterns they exhibit using various differential stain techniques that involve growing cells for two rounds of DNA replication in the presence of a base analog. In this example, regions of sister chromatids stained blue have one strand of the DNA labeled with a base analog while regions of sister chromatids stained green/yellow have both strands of the DNA labeled with a base analog.

The mechanisms of exchange between nonhomologous chromosomes and between sister chromatids may prove to share common features because the frequency of both events increases substantially in individuals with certain genetic disorders. These findings suggest that further study of sister chromatid exchange may contribute to the understanding of recombination mechanisms and to the relative stability of normal and genetically abnormal chromosomes. We shall encounter still another demonstration of SCEs when we consider replication of DNA (see Chapter 11).



EXPLORING GENOMICS

Mastering Genetics Visit the Study Area: Exploring Genomics

Human Chromosome Maps on the Internet

In this chapter we discussed how recombination data can be analyzed to develop chromosome maps based on linkage. Increasingly, chromosome maps are being developed using genomics techniques. As a result of the Human Genome Project, maps of human chromosomes are freely available on the Internet. In this exercise we explore the **National Center for Biotechnology Information (NCBI) Genes and Disease** web site to learn more about human chromosome maps.

NCBI Genes and Disease

Here we explore the Genes and Disease site, which presents human chromosome maps that show the locations of specific disease genes.

1. Access the Genes and Disease site at http://www.ncbi.nlm.nih.gov/books/NBK22183/
2. Under contents, click on "Chromosome Map" to see a page with an image of a karyotype of human chromosomes. Click on a chromosome in the chromosome map image, scroll down the page to view a chromosome, or click on a chromosome listed on the right side of the page. For example, click on chromosome 7. Notice that the number of genes on the chromosomes and the number of base pairs the chromosome contains are displayed above the image.
3. Look again at chromosome 7. At first you might think there are only five disease genes on this chromosome because the initial view shows only selected disease genes. However, if you click the "MapViewer" link for chromosome 7, you will see detailed information about the chromosome, including a complete "Master Map" of the genes it contains and the symbols used in naming genes.

   Explore features of this view, and be sure to look at the "Links" column, which provides access to OMIM (Online Mendelian Inheritance in Man, discussed in the Exploring Genomics feature for Chapter 3) data for a particular gene, as well as to protein information (*pr*) and lists of homologous genes (*hm;* these are other genes that have similar sequences).
4. Scan the chromosome maps in Map Viewer until you see one of the genes listed as a "hypothetical gene or protein."

   a. What does it mean if a gene or protein is referred to as hypothetical?

   b. What information do you think genome scientists use to assign a gene locus for a gene encoding a hypothetical protein?

Visit the **NCBI Map Viewer** homepage (http://www.ncbi.nlm.nih.gov/projects/mapview/) for an excellent database containing chromosome maps for a wide variety of different organisms.

CASE STUDY Links to autism

As parents of an autistic child, a couple decided that entering a research study would not only educate them about their child's condition but also help further research into this complex, behaviorally defined disorder. Researchers explained to the parents that autism results from the action of hundreds of genes as well as nongenetic factors. Modern DNA analysis techniques, including mapping studies, have identified a set of 9–18 genes with a high likelihood of involvement, referred to as candidate genes. Probing the genome at a deeper level has revealed as many as 2500 genes that might be risk factors. Generally unaware of the principles of basic genetics, the couple wondered if any future children they might have would be at risk of having autism and if prenatal diagnosis for autism is possible. The interviewer explained that if one child has autism, there is an approximate 25 percent risk of a future child having this condition. A prenatal test for autism is possible, but because the condition involves a potentially large number of genes and environmental conditions, such tests can only estimate, without any certainty, the likelihood of a second autistic child.

1. In a family with one autistic child the risk for another affected child is approximately 25 percent. This is the same level of risk that a couple who are each heterozygous for a recessive allele will have an affected child. What are the similarities and differences in these two situations?
2. Given that the prenatal test can provide only a probability estimate that the fetus will develop autism, what ethical issues should be discussed with the parents?

See Cox, D. (2017). Are we ready for a prenatal test for autism? (https://www.theguardian.com/science/blog/2014/may/01/prenatal-screening-test-autism-ethical-implications).

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Genes located on the same chromosome are said to be linked. Alleles of linked genes located close together on the same homolog are usually transmitted together during gamete formation.
2. Crossover frequency between linked genes during gamete formation is proportional to the distance between genes, providing the experimental basis for mapping the location of genes relative to one another along the chromosome.
3. Determining the sequence of genes in a three-point mapping experiment requires analysis of double-crossover gametes, as reflected in the phenotype of the offspring receiving those gametes.
4. Interference describes the extent to which a crossover in one region of a chromosome influences the occurrence of a crossover in an adjacent region of the chromosome and is quantified by calculating the coefficient of coincidence (C).
5. Human linkage studies, initially relying on pedigree and lod score analysis, and subsequently on somatic cell hybridization techniques, are now enhanced by the use of newly discovered molecular DNA markers.
6. Cytological investigations of both maize and *Drosophila* reveal that crossing over involves a physical exchange of segments between nonsister chromatids.
7. Recombination events are known to occur between sister chromatids in mitosis and are referred to as sister chromatid exchanges (SCEs).

INSIGHTS AND SOLUTIONS

1. In a series of two-point mapping crosses involving three genes linked on chromosome III in *Drosophila*, the following distances were calculated:

$$\mathit{cd}-\mathit{sr}\ 13\ \text{mu}$$
$$\mathit{cd}-\mathit{ro}\ 16\ \text{mu}$$

(a) Why can't we determine the sequence and construct a map of these three genes?

(b) What mapping data will resolve the issue?

(c) Can we tell which of the sequences shown here is correct?

$$\mathit{ro}\overset{16}{\text{——}}\mathit{cd}\overset{13}{\text{——}}\mathit{sr}$$

or

$$\mathit{sr}\overset{13}{\text{——}}\mathit{cd}\overset{16}{\text{——}}\mathit{ro}$$

Solution:

(a) It is impossible to do so because there are two possibilities based on these limited data:

Case 1: $\mathit{cd}\overset{13}{\text{——}}\mathit{sr}\overset{3}{\text{——}}\mathit{ro}$

or

Case 2: $\mathit{ro}\overset{16}{\text{——}}\mathit{cd}\overset{13}{\text{——}}\mathit{sr}$

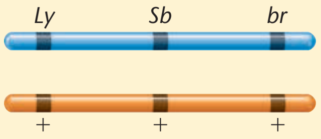
(b) The map distance is determined by crossing over between *ro* and *sr*. If case 1 is correct, it should be 3 mu, and if case 2 is correct, it should be 29 mu. In fact, this distance is 29 mu, demonstrating that case 2 is correct.

(c) No; based on the mapping data, they are equivalent.

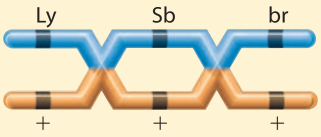
2. In *Drosophila, Lyra (Ly)* and *Stubble (Sb)* are dominant mutations located at loci 40 and 58, respectively, on chromosome III. A recessive mutation with bright red eyes was discovered and shown also to be on chromosome III. A map is obtained by crossing a female who is heterozygous for all three mutations to a male homozygous for the *bright red* mutation (which we refer to here as *br*). The data in the table are generated. Determine the location of the *br* mutation on chromosome III. By referring to Figure 5.13, predict what mutation has been discovered. How could you be sure?

Phenotype				Number
1	*Ly*	*Sb*	*br*	404
2	+	+	+	422
3	*Ly*	+	+	18
4	+	*Sb*	*br*	16
5	*Ly*	+	*br*	75
6	+	*Sb*	+	59
7	*Ly*	*Sb*	+	4
8	+	+	*br*	2
Total				1000

Solution: First determine the distribution of the alleles between the homologs of the heterozygous crossover parent (the female in this case). To do this, locate the most frequent reciprocal phenotypes, which arise from the noncrossover gametes. These are phenotypes 1 and 2. Each phenotype represents the alleles on one of the homologs. Therefore, the distribution of alleles is



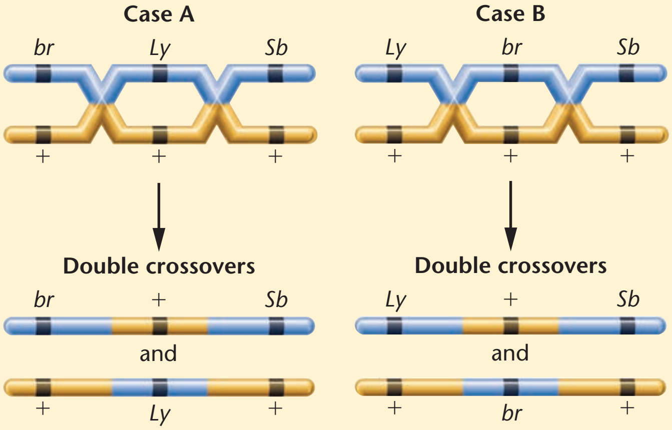
Second, determine the correct *sequence* of the three loci along the chromosome. This is done by determining which sequence yields the observed double-crossover phenotypes that are the least frequent reciprocal phenotypes (7 and 8). If the sequence is correct as written, then a double crossover, depicted here,



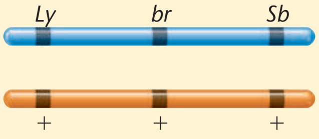
(continued)

Insights and Solutions—continued

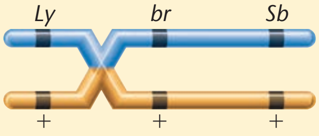
would yield *Ly* + *br* and + *Sb* + as phenotypes. Inspection shows that these categories (5 and 6) are actually single crossovers, not double crossovers. Therefore, the sequence, as written, is incorrect. There are only two other possible sequences. Either the *Ly* gene (case A) or the *br* gene (case B) is in the middle between the other two genes.



Comparison with the actual data shows that case B is correct. The double-crossover gametes 7 and 8 yield flies that express *Ly* and *Sb* but not *br, or* express *br* but not *Ly* and *Sb.* Therefore, the correct *arrangement* and *sequence* are as follows:



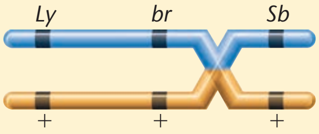
Once the sequence is found, determine the location of *br* relative to *Ly* and *Sb.* A single crossover between *Ly* and *br*, as shown here,



yields flies that are *Ly* + + and + *br Sb* (phenotypes 3 and 4). Therefore, the distance between the *Ly* and *br* loci is equal to

$$\frac{18 + 16 + 4 + 2}{1000} = \frac{40}{1000} = 0.04 = 4 \text{ mu}$$

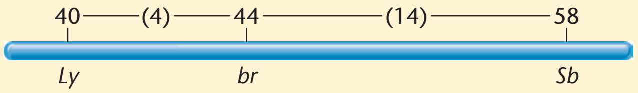
Remember that, because we need to know the frequency of all crossovers between *Ly* and *br*, we must add in the double crossovers, since they represent two single crossovers occurring simultaneously. Similarly, the distance between the *br* and *Sb* loci is derived mainly from single crossovers between them.



This event yields *Ly br* + and + + *Sb* phenotypes (phenotypes 5 and 6). Therefore, the distance equals

$$\frac{75 + 59 + 4 + 2}{1000} = \frac{140}{1000} = 0.14 = 14 \text{ mu}$$

The final map shows that *br* is located at locus 44, since *Lyra* and *Stubble* are known:



Inspection of Figure 5.13 reveals that the mutation *scarlet*, which produces bright red eyes, is known to sit at locus 44, so it is reasonable to hypothesize that the *bright red* eye mutation is an allele of *scarlet.* To test this hypothesis, we could cross females of our *bright red* mutant with known *scarlet* males. If the two mutations are alleles, no complementation will occur, and all progeny will reveal a bright red mutant eye phenotype. If complementation occurs, all progeny will show normal brick-red (wild-type) eyes, since the *bright red* mutation and *scarlet* are at different loci. (They are probably very close together.) In such a case, all progeny will be heterozygous at both the *bright red* and the *scarlet* loci and will not express either mutation because they are both recessive. This cross represents what is called an *allelism test.*

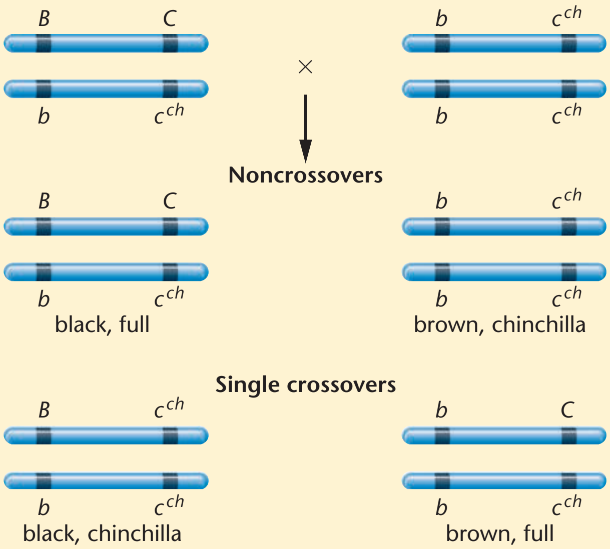
3. In rabbits, black color (*B*) is dominant to brown (*b*), while full color (*C*) is dominant to chinchilla (c^{ch}). The genes controlling these traits are linked. Rabbits that are heterozygous for both traits and express black, full color were crossed with rabbits that express brown, chinchilla, with the following results:

   31 brown, chinchilla
   34 black, full color
   16 brown, full color
   19 black, chinchilla

   Determine the arrangement of alleles in the heterozygous parents and the map distance between the two genes.

Solution: This is a two-point mapping problem. The two most prevalent reciprocal phenotypes are the noncrossovers, and the less frequent reciprocal phenotypes arise from a single crossover. The distribution of alleles is derived from the noncrossover phenotypes because they enter gametes intact.

The single crossovers give rise to 35/100 offspring (35 percent). Therefore, the distance between the two genes is 35 mu.



Problems and Discussion Questions

1. **HOW DO WE KNOW?** In this chapter, we focused on linkage, chromosomal mapping, and many associated phenomena. In the process, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?

(a) How was it established experimentally that the frequency of recombination (crossing over) between two genes is related to the distance between them along the chromosome?

(b) How do we know that specific genes are linked on a single chromosome, in contrast to being located on separate chromosomes?

(c) How do we know that crossing over results from a physical exchange between chromatids?

(d) How do we know that sister chromatids undergo recombination during mitosis?

(e) When designed matings cannot be conducted in an organism (for example, in humans), how do we learn that genes are linked, and how do we map them?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 94. Most of these center around the process of crossing over between linked genes. Write a short essay that discusses how crossing over can be detected and how the resultant data provide the basis of chromosome mapping.

3. Describe the cytological observation that suggests that crossing over occurs during the first meiotic prophase.

4. Why does more crossing over occur between two distantly linked genes than between two genes that are very close together on the same chromosome?

5. Explain why a 50 percent recovery of single-crossover products is the upper limit, even when crossing over *always* occurs between two linked genes?

6. Why are double-crossover events expected less frequently than single-crossover events?

7. What is the proposed basis for positive interference?

8. What two essential criteria must be met in order to execute a successful mapping cross?

9. The genes *dumpy* (*dp*), *clot* (*cl*), and *apterous* (*ap*) are linked on chromosome II of *Drosophila*. In a series of two-point mapping crosses, the following genetic distances were determined. What is the sequence of the three genes?

dp—*ap*	42
dp—*cl*	3
ap—*cl*	39

10. Colored aleurone in the kernels of corn is due to the dominant allele *R*. The recessive allele *r*, when homozygous, produces colorless aleurone. The plant color (not the kernel color) is controlled by another gene with two alleles, *Y* and *y*. The dominant *Y* allele results in green color, whereas the homozygous presence of the recessive *y* allele causes the plant to appear yellow. In a testcross between a plant of unknown genotype and phenotype and a plant that is homozygous recessive for both traits, the following progeny were obtained:

colored, green	88
colored, yellow	12
colorless, green	8
colorless, yellow	92

Explain how these results were obtained by determining the exact genotype and phenotype of the unknown plant, including the precise arrangement of the alleles on the homologs.

11. In the cross shown here, involving two linked genes, *ebony* (*e*) and *claret* (*ca*), in *Drosophila*, where crossing over does not occur in males, offspring were produced in a 2 + :1 *ca* : 1 *e* phenotypic ratio:

$$\overset{\text{♀}}{\frac{e \quad ca^{+}}{e^{+} \quad ca}} \times \overset{\text{♂}}{\frac{e \quad ca^{+}}{e^{+} \quad ca}}$$

These genes are 30 units apart on chromosome III. What did crossing over in the female contribute to these phenotypes?

12. In a series of two-point mapping crosses involving five genes located on chromosome II in *Drosophila*, the following recombinant (single-crossover) frequencies were observed:

pr—*adp*	29%
pr—*vg*	13
pr—*c*	21
pr—*b*	6
adp—*b*	35
adp—*c*	8
adp—*vg*	16
vg—*b*	19
vg—*c*	8
c—*b*	27

(a) Given that the *adp* gene is near the end of chromosome II (locus 83), construct a map of these genes.

(b) In another set of experiments, a sixth gene, *d*, was tested against *b* and *pr:*

d—*b*	17%
d—*pr*	23%

Predict the results of two-point mapping between *d* and *c*, *d* and *vg*, and *d* and *adp*.

13. Two different female *Drosophila* were isolated, each heterozygous for the autosomally linked genes *b* (*black body*), *d* (*dachs tarsus*), and *c* (*curved wings*). These genes are in the order *d*—*b*—*c*, with *b* being closer to *d* than to *c*. Shown here is the genotypic arrangement for each female along with the various gametes formed by both:

Female A		Female B	
$\dfrac{d\,b\,+}{+\,+\,c}$		$\dfrac{d\,+\,+}{+\,b\,c}$	
↓	**Gamete formation**	↓	
(1) d b c	(5) d + +	(1) d b +	(5) d b c
(2) + + +	(6) + b c	(2) + + c	(6) + + +
(3) + + c	(7) d + c	(3) d + c	(7) d + +
(4) d b +	(8) + b +	(4) + b +	(8) + b C

Identify which categories are noncrossovers (NCOs), single crossovers (SCOs), and double crossovers (DCOs) in each case. Then, indicate the relative frequency in which each will be produced.

14. In *Drosophila*, a cross was made between females—all expressing the three X-linked recessive traits *scute* bristles (*sc*), *sable* body (*s*), and *vermilion* eyes (*v*)—and wild-type males. In the F_1, all females were wild type, while all males expressed all three mutant traits. The cross was carried to the F_2 generation, and 1000 offspring were counted, with the results shown in the following table.

Phenotype	Offspring
sc *s* *v*	314
+ + +	280
+ *s* *v*	150
sc + +	156
sc + *v*	46
+ *s* +	30
sc *s* +	10
+ + *v*	14

No determination of sex was made in the data.

(a) Using proper nomenclature, determine the genotypes of the P_1 and F_1 parents.
(b) Determine the sequence of the three genes and the map distances between them.
(c) Are there more or fewer double crossovers than expected?
(d) Calculate the coefficient of coincidence. Does it represent positive or negative interference?

15. Another cross in *Drosophila* involved the recessive, X-linked genes *yellow* (*y*), *white* (*w*), and *cut* (*ct*). A yellow-bodied, white-eyed female with normal wings was crossed to a male whose eyes and body were normal but whose wings were cut. The F_1 females were wild type for all three traits, while the F_1 males expressed the yellow-body and white-eye traits. The cross was carried to an F_2 progeny, and only male offspring were tallied. On the basis of the data shown here, a genetic map was constructed.

Phenotype	Male Offspring
y + *ct*	9
+ *w* +	6
y *w* *ct*	90
+ + +	95
+ + *ct*	424
y *w* +	376
y + +	0
+ *w* *ct*	0

(a) Diagram the genotypes of the F_1 parents.
(b) Construct a map, assuming that *white* is at locus 1.5 on the X chromosome.
(c) Were any double-crossover offspring expected?
(d) Could the F_2 female offspring be used to construct the map? Why or why not?

16. In *Drosophila, Dichaete* (*D*) is a mutation on chromosome III with a dominant effect on wing shape. It is lethal when homozygous. The genes *ebony* body (*e*) and *pink* eye (*p*) are recessive mutations on chromosome III. Flies from a *Dichaete* stock were crossed to homozygous ebony, pink flies, and the F_1 progeny, with a Dichaete phenotype, were backcrossed to the ebony, pink homozygotes. Using the results of this backcross shown in the table,
(a) Diagram this cross, showing the genotypes of the parents and offspring of both crosses.
(b) What is the sequence and interlocus distance between these three genes?

Phenotype	Number
Dichaete	401
ebony, pink	389
Dichaete, ebony	84
pink	96
Dichaete, pink	2
ebony	3
Dichaete, ebony, pink	12
wild type	13

17. *Drosophila* females homozygous for the third chromosomal genes *pink* and *ebony* (the same genes from Problem 16) were crossed with males homozygous for the second chromosomal gene *dumpy*. Because these genes are recessive, all offspring were wild type (normal). F_1 females were testcrossed to triply recessive males. If we assume that the two linked genes, *pink* and *ebony*, are 20 mu apart, predict the results of this cross. If the reciprocal cross were made (F_1 males—where no crossing over occurs—with triply recessive females), how would the results vary, if at all?

18. In *Drosophila*, two mutations, *Stubble* (*Sb*) and *curled* (*cu*), are linked on chromosome III. *Stubble* is a dominant gene that is lethal in a homozygous state, and *curled* is a recessive gene. If a female of the genotype

$$\frac{Sb\quad cu}{+\quad +}$$

is to be mated to detect recombinants among her offspring, what male genotype would you choose as a mate?

19. If the cross described in Problem 18 were made, and if *Sb* and *cu* are 8.2 map units apart on chromosome III, and if 1000 offspring were recovered, what would be the outcome of the cross, assuming that equal numbers of males and females were observed?

20. Are mitotic recombinations and sister chromatid exchanges effective in producing genetic variability in an individual? in the offspring of individuals?

21. What possible conclusions can be drawn from the observations that in male *Drosophila*, no crossing over occurs, and that during meiosis, synaptonemal complexes are not seen in males but are observed in females where crossing over occurs?

22. An organism of the genotype *AaBbCc* was testcrossed to a triply recessive organism (*aabbcc*). The genotypes of the progeny are presented in the following table.

20	*AaBbCc*	20	*AaBbcc*
20	*aabbCc*	20	*aabbcc*
5	*AabbCc*	5	*Aabbcc*
5	*aaBbCc*	5	*aaBbcc*

(a) If these three genes were all assorting independently, how many genotypic and phenotypic classes would result in the offspring, and in what proportion, assuming simple dominance and recessiveness in each gene pair?
(b) Answer part (a) again, assuming the three genes are so tightly linked on a single chromosome that no crossover gametes were recovered in the sample of offspring.
(c) What can you conclude from the *actual* data about the location of the three genes in relation to one another?

23. Based on our discussion of the potential inaccuracy of mapping (see Figure 5.12), would you revise your answer to Problem 22? If so, how?

24. Traditional gene mapping has been applied successfully to a variety of organisms including yeast, fungi, maize, and *Drosophila*. However, human gene mapping has only recently shared a similar spotlight. What factors have delayed the application of traditional gene-mapping techniques in humans?

25. DNA markers have greatly enhanced the mapping of genes in humans. What are DNA markers, and what advantage do they confer?

26. In a certain plant, fruit is either red or yellow, and fruit shape is either oval or long. Red and oval are the dominant traits. Two plants, both heterozygous for these traits, were testcrossed, with the following results.

	Progeny	
Phenotype	**Plant A**	**Plant B**
red, long	46	4
yellow, oval	44	6
red, oval	5	43
yellow, long	5	47
	100	100

Determine the location of the genes relative to one another and the genotypes of the two parental plants.

27. Two plants in a cross were each heterozygous for two gene pairs (*Ab*/*aB*) whose loci are linked and 25 mu apart. Assuming that crossing over occurs during the formation of both male and female gametes and that the *A* and *B* alleles are dominant, determine the phenotypic ratio of their offspring.

Extra-Spicy Problems

28. A number of human–mouse somatic cell hybrid clones were examined for the expression of specific human genes and the presence of human chromosomes. The results are summarized in the following table. Assign each gene to the chromosome on which it is located.

	Hybrid Cell Clone					
	A	B	C	D	E	F
Genes expressed						
ENO1 (enolase-1)	−	+	−	+	+	−
MDH1 (malate dehydrogenase-1)	+	+	−	+	−	+
PEPS (peptidase S)	+	−	+	−	−	−
PGM1 (phosphoglucomutase-1)	−	+	−	+	+	−
Chromosomes (present or absent)						
1	−	+	−	+	+	−
2	+	+	−	+	−	+
3	+	+	−	−	+	−
4	+	−	+	−	−	−
5	−	+	+	+	+	+

29. A female of genotype

$$\frac{a \quad b \quad c}{+ \quad + \quad +}$$

produces 100 meiotic tetrads. Of these, 68 show no crossover events. Of the remaining 32, 20 show a crossover between *a* and *b*, 10 show a crossover between *b* and *c*, and 2 show a double crossover between *a* and *b* and between *b* and *c*. Of the 400 gametes produced, how many of each of the 8 different genotypes will be produced? Assuming the order *a*–*b*–*c* and the allele arrangement previously shown, what is the map distance between these loci?

30. In laboratory class, a genetics student was assigned to study an unknown mutation in *Drosophila* that had a whitish eye. He crossed females from his true-breeding mutant stock to wild-type (brick-red-eyed) males, recovering all wild-type F_1 flies. In the F_2 generation, the following offspring were recovered in the following proportions:

wild type	5/8
bright red	1/8
brown eye	1/8
white eye	1/8

The student was stumped until the instructor suggested that perhaps the whitish eye in the original stock was the result of homozygosity for a mutation causing brown eyes *and* a mutation causing bright red eyes, illustrating gene interaction (see Chapter 4). After much thought, the student was able to analyze the data, explain the results, and learn several things about the location of the two genes relative to one another. One key to his understanding was that crossing over occurs in *Drosophila* females but not in males. Based on his analysis, what did the student learn about the two genes?

31. *Drosophila melanogaster* has one pair of sex chromosomes (XX or XY) and three pairs of autosomes, referred to as chromosomes II, III, and IV. A genetics student discovered a male fly with very short (*sh*) legs. Using this male, the student was able to establish a pure breeding stock of this mutant and found that it was recessive. She then incorporated the mutant into a stock containing the recessive gene *black* (*b*, body color located on chromosome II) and the recessive gene *pink* (*p*, eye color located on chromosome III). A female from the homozygous black, pink, short stock was then mated to a wild-type male. The F_1 males of this cross were all wild type and were then backcrossed to the homozygous *b, p, sh* females. The F_2 results appeared as shown in the following table. No other phenotypes were observed.

	Wild	Pink*	Black, Short*	Black, Pink, Short
Females	63	58	55	69
Males	59	65	51	60

*Other trait or traits are wild type.

(a) Based on these results, the student was able to assign *short* to a linkage group (a chromosome). Which one was it? Include your step-by-step reasoning.

(b) The student repeated the experiment, making the reciprocal cross, F_1 females backcrossed to homozygous *b, p, sh* males. She observed that 85 percent of the offspring fell into the given classes, but that 15 percent of the offspring were equally divided among *b* + *p*, *b* + +, + *sh p*, and + *sh* + phenotypic males and females. How can these results be explained, and what information can be derived from the data?

32. In *Drosophila*, a female fly is heterozygous for three mutations, *Bar* eyes (*B*), *miniature* wings (*m*), and *ebony* body (*e*). Note that *Bar* is a dominant mutation. The fly is crossed to a male with normal eyes, miniature wings, and ebony body. The results of the cross are as follows.

111 miniature	101 Bar, ebony
29 wild type	31 Bar, miniature, ebony
117 Bar	35 ebony
26 Bar, miniature	115 miniature, ebony

Interpret the results of this cross. If you conclude that linkage is involved between any of the genes, determine the map distance(s) between them.

33. The gene controlling the Xg blood group alleles (Xg^+ and Xg^-) and the gene controlling a newly described form of inherited recessive muscle weakness called *episodic muscle weakness* (*EMWX*) (Ryan et al., 1999) are closely linked on the X chromosome in humans at position Xp22.3 (the tip of the short arm). A male with EMWX who is Xg^- marries a woman who is Xg^+, and they have eight daughters and one son, all of whom are normal for muscle function, the male being Xg^+ and all the daughters being heterozygous at both the *EMWX* and *Xg* loci. Following is a table that lists three of the daughters with the phenotypes of their husbands and children.

	Husband's Phenotype	Offspring's Sex	Offspring's Phenotype
Daughter 1:	Xg^+	female	Xg^+
		male	EMWX, Xg^+
Daughter 2:	Xg^-	male	Xg^-
		female	Xg^+
		male	EMWX, Xg^-
Daughter 3:	Xg^-	male	EMWX, Xg^-
		male	Xg^+
		male	Xg^-
		male	EMWX, Xg^+
		male	Xg^-
		male	EMWX, Xg^-
		female	Xg^+
		female	Xg^-
		female	Xg^+

(a) Create a pedigree that represents all data stated above and in the following table.

(b) For each of the offspring, indicate whether or not a crossover was required to produce the phenotypes that are given.

34. Because of the relatively high frequency of meiotic errors that lead to developmental abnormalities in humans, many research efforts have focused on identifying correlations between error frequency and chromosome morphology and behavior. Tease et al. (2002) studied human fetal oocytes of chromosomes 21, 18, and 13 using an immunocytological approach that allowed a direct estimate of the frequency and position of meiotic recombination. Below is a summary of information [modified from Tease et al. (2002)] that compares recombination frequency with the frequency of trisomy for chromosomes 21, 18, and 13. (*Note:* You may want to read appropriate portions of Chapter 8 for descriptions of these trisomic conditions.)

Trisomic	Mean Recombination Frequency	Live-born Frequency
Chromosome 21	1.23	1/700
Chromosome 18	2.36	1/3000–1/8000
Chromosome 13	2.50	1/5000–1/19,000

(a) What conclusions can be drawn from these data in terms of recombination and nondisjunction frequencies? How might recombination frequencies influence trisomic frequencies?

(b) Other studies indicate that the number of crossovers per oocyte is somewhat constant, and it has been suggested that positive chromosomal interference acts to spread out a limited number of crossovers among as many chromosomes as possible. Considering information in part (a), speculate on the selective advantage positive chromosomal interference might confer.

6

Transmission electron micrograph of conjugating *Escherichia coli*.

Genetic Analysis and Mapping in Bacteria and Bacteriophages

CHAPTER CONCEPTS

- Bacterial genomes are most often contained in a single circular chromosome.
- Bacteria have developed numerous ways of exchanging and recombining genetic information between individual cells, including conjugation, transformation, and transduction.
- The ability to undergo conjugation and to transfer the bacterial chromosome from one cell to another is governed by genetic information contained in the DNA of a "fertility," or F, factor.
- The F factor can exist autonomously in the bacterial cytoplasm as a plasmid, or it can integrate into the bacterial chromosome, where it facilitates the transfer of the host chromosome to the recipient cell, leading to genetic recombination.
- Genetic recombination during conjugation provides the basis for mapping bacterial genes.
- Bacteriophages are viruses that have bacteria as their hosts. Viral DNA is injected into the host cell, where it replicates and directs the reproduction of the bacteriophage and the lysis of the bacterium.
- Rarely, following infection, bacteriophage DNA integrates into the host chromosome, becoming a prophage, where it is replicated along with the bacterial DNA.
- Bacteriophages undergo both intergenic and intragenic recombination.

In this chapter, we shift from consideration of transmission genetics and mapping in eukaryotes to discussion of the analysis of genetic recombination and mapping in bacteria and bacteriophages, viruses that have bacteria as their host. As we focus on these topics, it will become clear that complex processes have evolved in bacteria and bacteriophages that transfer genetic information between individual cells within populations. These processes provide geneticists with the basis for chromosome mapping.

The study of bacteria and bacteriophages has been essential to the accumulation of knowledge in many areas of genetic study. For example, much of what we know about the expression and regulation of genetic information was initially derived from experimental work with them. Furthermore, as we shall see (Chapter 20), our extensive knowledge of bacteria and their resident plasmids has served as the basis for their widespread use in DNA cloning and other recombinant DNA procedures.

The value of bacteria and their viruses as research organisms in genetics is based on two important characteristics that they display. First, they have extremely short reproductive cycles. Literally hundreds of generations, amounting to billions of genetically identical bacteria or phages, can be produced in short periods of time. Second, they can also be studied in pure culture. That is, a single species or mutant strain of bacteria or one type of virus can with ease be isolated and investigated independently of other similar organisms. As a result, they have been indispensable to the progress made in genetics over the past half century.

6.1 Bacteria Mutate Spontaneously and Grow at an Exponential Rate

Genetic studies using bacteria depend on our ability to study mutations in these organisms. It has long been known that genetically homogeneous cultures of bacteria occasionally give rise to cells exhibiting heritable variation, particularly with respect to growth under unique environmental conditions. Prior to 1943, the source of this variation was hotly debated. The majority of bacteriologists believed that environmental factors induced changes in certain bacteria, leading to their survival or adaptation to the new conditions. For example, strains of *E. coli* are known to be sensitive to infection by the bacteriophage T1. Infection by the bacteriophage T1 leads to reproduction of the virus at the expense of the bacterial host, from which new phages are released as the host cell is disrupted, or lysed. If a plate of *E. coli* is uniformly sprayed with T1, almost all cells are lysed. Rare *E. coli* cells, however, survive infection and are not lysed. If these cells are isolated and established in pure culture, all their descendants are resistant to T1 infection. The **adaptation hypothesis,** put forth to explain this type of observation, implies that the interaction of the phage and bacterium is essential to the acquisition of immunity. In other words, exposure to the phage “induces” resistance in the bacteria.

On the other hand, the occurrence of **spontaneous mutations,** which occur regardless of the presence or absence of bacteriophage T1, suggested an alternative model to explain the origin of resistance in *E. coli*. In 1943, Salvador Luria and Max Delbrück presented the first convincing evidence that bacteria, like eukaryotic organisms, are capable of spontaneous mutation. Their experiment, referred to as the **fluctuation test,** marks the initiation of modern bacterial genetic study. (We will explore this discovery in Chapter 16.) Mutant cells that arise spontaneously in otherwise pure cultures can be isolated and established independently from the parent strain by the use of selection techniques. *Selection* refers to culturing the organism under conditions where only the desired mutant grows well, while the wild type does not grow. With carefully designed selection, mutations for almost any desired characteristic can now be isolated. Because bacteria and viruses usually contain only one copy of a single chromosome, and are therefore haploid, all mutations are expressed directly in the descendants of mutant cells, adding to the ease with which these microorganisms can be studied.

Bacteria are grown either in a liquid culture medium or in a petri dish on a semisolid agar surface. If the nutrient components of the growth medium are very simple and consist only of an organic carbon source (such as glucose or lactose) and various inorganic ions, including Na^+, K^+, Mg^{2+}, Ca^{2+}, and NH_4^+ present as inorganic salts, it is called **minimal medium.** To grow on such a medium, a bacterium must be able to synthesize all essential organic compounds (e.g., amino acids, purines, pyrimidines, sugars, vitamins, and fatty acids). A bacterium that can accomplish this remarkable biosynthetic feat—one that the human body cannot duplicate—is a **prototroph.** It is said to be wild type for all growth requirements. On the other hand, if a bacterium loses, through mutation, the ability to synthesize one or more organic components, it is an **auxotroph.** For example, a bacterium that loses the ability to make histidine is designated as a *his*$^-$ auxotroph, in contrast to its prototrophic *his*$^+$ counterpart. For the *his*$^-$ bacterium to grow, this amino acid must be added as a supplement to the minimal medium. Medium that has been extensively supplemented is called *complete medium*.

To study bacterial growth quantitatively, an inoculum of bacteria—a small amount of a bacteria-containing solution, for example, 0.1 or 1.0 mL—is placed in liquid culture medium. A graph of the characteristic growth pattern for a bacteria culture is shown in **Figure 6.1**. Initially, during the **lag phase,** growth is slow. Then, a period of rapid growth, called the **logarithmic (log) phase,** ensues. During this phase, cells divide continually with a fixed time interval between cell divisions, resulting in exponential growth. When a cell density of about 10^9 cells/mL

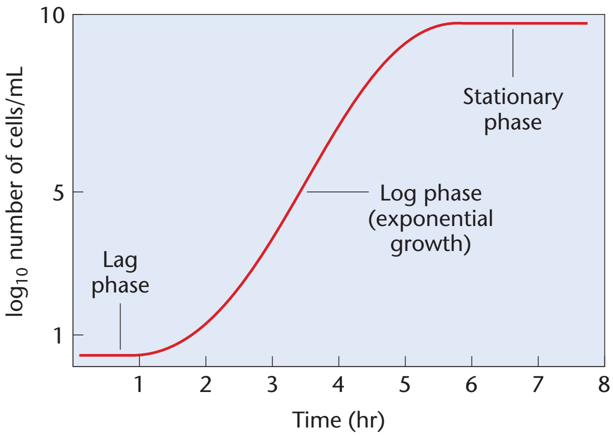


FIGURE 6.1 Typical bacterial population growth curve showing the initial lag phase, the subsequent log phase where exponential growth occurs, and the stationary phase that occurs when nutrients are exhausted. Eventually, all cells will die.

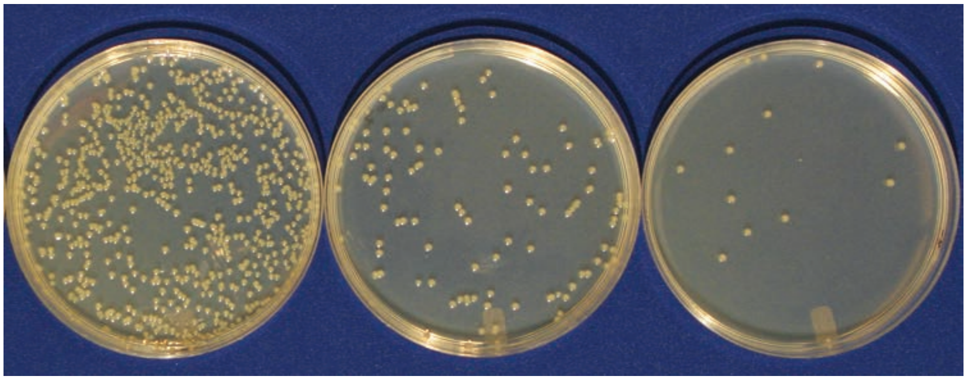


FIGURE 6.2 Results of the serial dilution technique and subsequent culture of bacteria. Each dilution varies by a factor of 10. Each colony was derived from a single bacterial cell.

of culture medium is reached, nutrients become limiting and cells cease dividing; at this point, the cells enter the **stationary phase.** The doubling time during the log phase can be as short as 20 minutes. Thus, an initial inoculum of a few thousand cells added to the culture easily achieves maximum cell density during an overnight incubation.

Cells grown in liquid medium can be quantified by transferring them to the semisolid medium of a petri dish. Following incubation and many divisions, each cell gives rise to a colony visible on the surface of the medium. By counting colonies, it is possible to estimate the number of bacteria present in the original culture. If the number of colonies is too great to count, then successive dilutions (in a technique called *serial dilution*) of the original liquid culture are made and plated, until the colony number is reduced to the point where it can be counted (Figure 6.2). This technique allows the number of bacteria present in the original culture to be calculated.

For example, let's assume that the three petri dishes in Figure 6.2 represent dilutions of the liquid culture by 10^{-3}, 10^{-4}, and 10^{-5} (from left to right).* We need only select the dish in which the number of colonies can be counted accurately. Because each colony presumably arose from a single bacterium, the number of colonies times the dilution factor represents the number of bacteria in each milliliter (mL) of the initial inoculum before it was diluted. In Figure 6.2, the rightmost dish contains 12 colonies. The dilution factor for a 10^{-5} dilution is 10^5. Therefore, the initial number of bacteria was 12×10^5 per mL.

6.2 Genetic Recombination Occurs in Bacteria

Development of techniques that allowed the identification and study of bacterial mutations led to detailed investigations of the transfer of genetic information between individual organisms. As we shall see, as with meiotic crossing over in eukaryotes, the process of **genetic recombination** in bacteria provided the basis for the development of chromosome mapping methodology. It is important to note at the outset of our discussion that the term *genetic recombination,* as applied to bacteria, refers to the *replacement* of one or more genes present in the chromosome of one cell with those from the chromosome of a genetically distinct cell. While this is somewhat different from our use of the term in eukaryotes—where it describes crossing over *resulting in a reciprocal exchange*—the overall effect is the same: Genetic information is transferred and results in an altered genotype.

We will discuss three processes that result in the transfer of genetic information from one bacterium to another: *conjugation, transformation,* and *transduction*. Collectively, knowledge of these processes has helped us understand the origin of genetic variation between members of the same bacterial species, and in some cases, between members of different species. When transfer of genetic information occurs between members of the same species, the term **vertical gene transfer** applies. When transfer occurs between members of related but distinct bacterial species, the term **horizontal gene transfer** is used. The horizontal gene transfer process has played a significant role in the evolution of bacteria. Often, the genes discovered to be involved in horizontal transfer are those that also confer survival advantages to the recipient species. For example, one species may transfer antibiotic resistance genes to another species. Or genes conferring enhanced pathogenicity may be transferred. Thus, the potential for such transfer is a major concern in the medical community. In addition, horizontal gene transfer has been a major factor in the process of speciation in bacteria. Many, if not most, bacterial species have been the recipient of genes from other species.

Conjugation in Bacteria: The Discovery of F^+ and F^- Strains

Studies of bacterial recombination began in 1946, when Joshua Lederberg and Edward Tatum showed that bacteria undergo **conjugation,** a process by which genetic information from one bacterium is transferred to and recombined with that of another bacterium. Their initial experiments

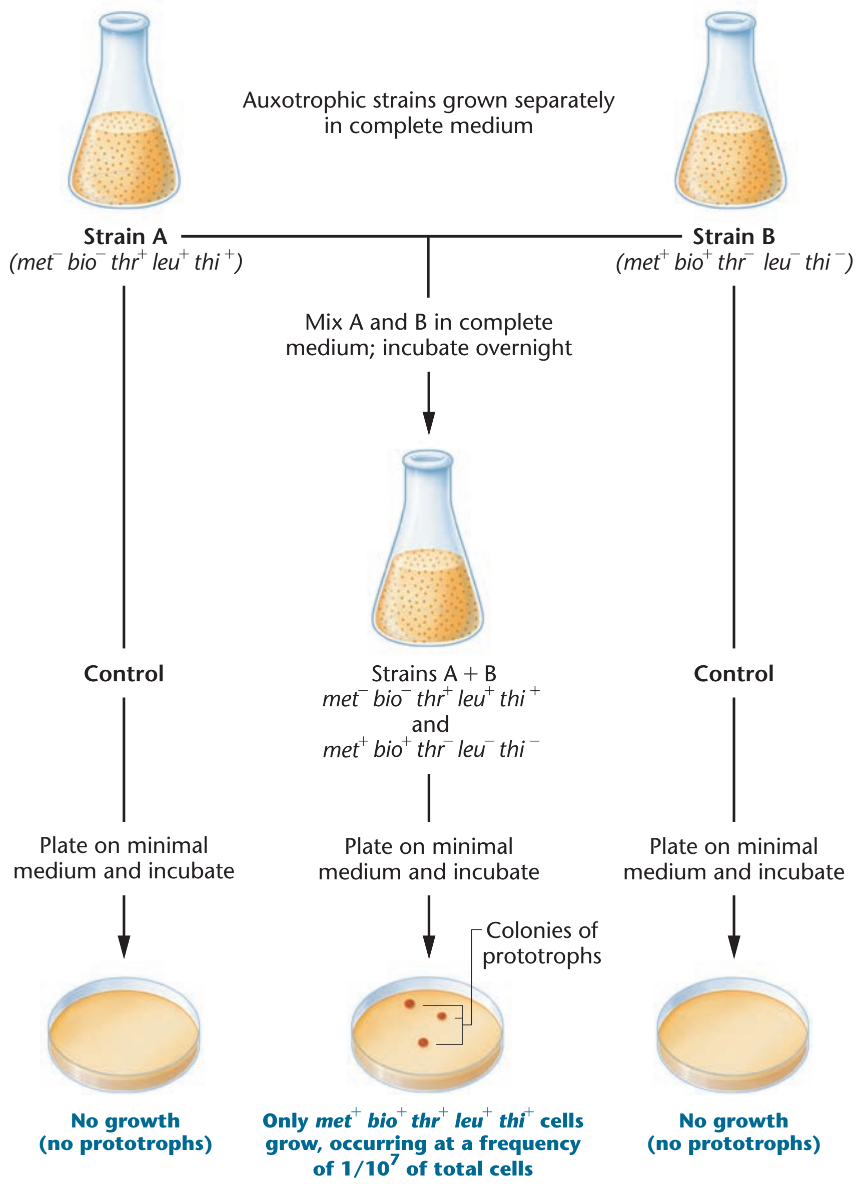
* 10^{-5} represents a 1:100,000 dilution.

were performed with two multiple auxotrophs (nutritional mutants) of *E. coli* strain K12. As shown in **Figure 6.3**, strain A required methionine (met) and biotin (bio) in order to grow, whereas strain B required threonine (thr), leucine (leu), and thiamine (thi). Neither strain would grow on minimal medium. The two strains were first grown separately in supplemented media, and then cells from both were mixed and grown together for several more generations. They were then plated on minimal medium. Any cells that grew on minimal medium were prototrophs. It is highly improbable that any of the cells containing two or three mutant genes would undergo **spontaneous mutation** simultaneously at two or three independent locations to become wild-type cells. Therefore, the researchers assumed that any prototrophs recovered must have arisen as a result of some form of genetic exchange and recombination between the two mutant strains.

In this experiment, prototrophs were recovered at a rate of $1/10^7$ (or 10^{-7}) cells plated. The controls for this experiment consisted of separate plating of cells from strains A and B on minimal medium. No prototrophs were recovered. On the basis of these observations, Lederberg and Tatum proposed that, while the events were indeed rare, genetic recombination had occurred.

Lederberg and Tatum's findings were soon followed by numerous experiments that elucidated the physical nature and the genetic basis of conjugation. It quickly became evident that different strains of bacteria were capable of effecting a unidirectional transfer of genetic material. When cells serve as donors of parts of their chromosomes, they are designated as **F^+ cells** (F for "fertility"). Recipient bacteria, designated as **F^- cells,** receive the donor chromosome material (now known to be DNA) and recombine it with part of their own chromosome.

Experimentation subsequently established that cell-to-cell contact is essential for chromosome transfer. Support for this concept was provided by Bernard Davis, who designed the Davis U-tube for growing F^+ and F^- cells (**Figure 6.4**). At the base of the tube is a sintered glass filter with a pore size that allows passage of the liquid medium but is too small to allow passage of bacteria. The F^+ cells are placed on one side of the filter and F^- cells on the other side. The medium passes back and forth across the filter so that it is shared by both sets of bacterial cells during incubation. When Davis plated samples from both sides of the tube on minimal medium, no prototrophs were found, and he logically concluded that *physical contact between cells of the two strains is essential to genetic recombination*. We now know that this physical interaction is the initial stage of the process of conjugation and is mediated by a structure called the **F pilus** (or **sex pilus;** pl. pili), a 6- to 9-nm tubular extension of the cell (see the chapter-opening photograph on p. 123). Bacteria often have many pili of different types performing different cellular functions, but all pili are involved in some way with adhesion (the binding together of cells).



Auxotrophic strains grown separately in complete medium
Strain A
($met^- bio^- thr^+ leu^+ thi^+$)
Strain B
($met^+ bio^+ thr^- leu^- thi^-$)
Mix A and B in complete medium; incubate overnight
Control
Strains A + B
$met^- bio^- thr^+ leu^+ thi^+$
and
$met^+ bio^+ thr^- leu^- thi^-$
Control
Plate on minimal medium and incubate
Plate on minimal medium and incubate
Plate on minimal medium and incubate
Colonies of prototrophs
No growth (no prototrophs)
Only $met^+ bio^+ thr^+ leu^+ thi^+$ cells grow, occurring at a frequency of $1/10^7$ of total cells
No growth (no prototrophs)

FIGURE 6.3 Production of prototrophs as a result of genetic recombination between two auxotrophic strains. Neither auxotrophic strain will grow on minimal medium, but prototrophs do, suggesting that genetic recombination has occurred.

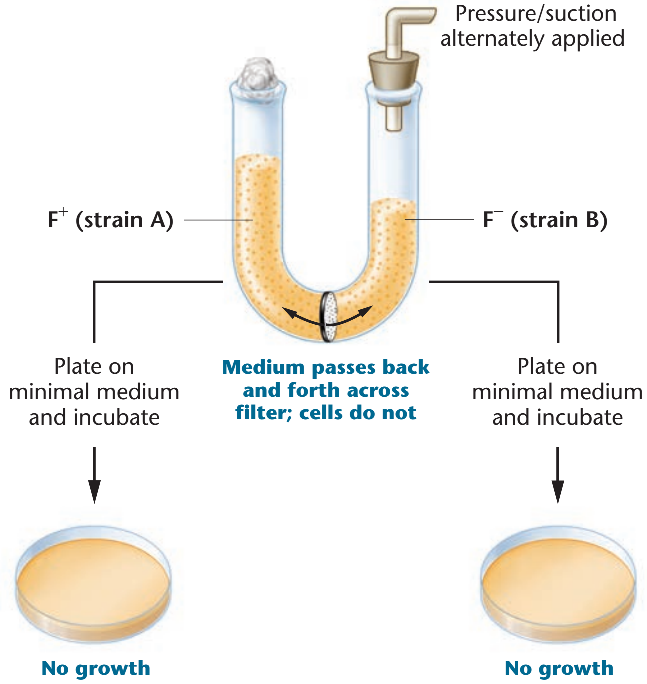

Pressure/suction alternately applied
F^+ (strain A)
F^- (strain B)
Plate on minimal medium and incubate
Medium passes back and forth across filter; cells do not
Plate on minimal medium and incubate
No growth
No growth

FIGURE 6.4 When strain A and strain B auxotrophs are grown in a common medium but separated by a filter, as in this Davis U-tube apparatus, no genetic recombination occurs and no prototrophs are produced.

After contact has been initiated between mating pairs, chromosome transfer is possible.

Later evidence established that F^+ cells contain a **fertility factor (F factor)** that confers the ability to donate part of their chromosome during conjugation. Experiments by Joshua and Esther Lederberg and by William Hayes and Luca Cavalli-Sforza showed that certain environmental conditions eliminate the F factor from otherwise fertile cells. However, if these "infertile" cells are then grown with fertile donor cells, the F factor is regained. These findings led to the hypothesis that the F factor is a mobile element, a conclusion further supported by the observation that, after conjugation, recipient F^- cells always become F^+. Thus, in addition to the rare cases of gene transfer (genetic recombination) that result from conjugation, the F factor itself is passed to *all* recipient cells. Accordingly, the initial cross of Lederberg and Tatum (Figure 6.3) can be described as follows:

Strain A		Strain B
F^+	×	F^-
(DONOR)		(RECIPIENT)

Characterization of the F factor confirmed these conclusions. Like the bacterial chromosome, though distinct from it, the F factor has been shown to consist of a circular, double-stranded DNA molecule; it is equivalent in size to about 2 percent of the bacterial chromosome (about 100,000 nucleotide pairs) and contains as many as 40 genes. Many are *tra* genes, whose products are involved in the *tra*nsfer of genetic information, including the genes essential to the formation of the sex pilus.

Geneticists believe that transfer of the F factor during conjugation involves separation of the two strands of its double helix and movement of one of the two strands into the recipient cell. The other strand remains in the donor cell. Both strands, one moving across the conjugation tube and one remaining in the donor cell, are replicated. The result is that both the donor and the recipient cells become F^+. This process is diagrammed in Figure 6.5.

To summarize, an *E. coli* cell may or may not contain the F factor. When it is present, the cell is able to form a sex pilus and potentially serve as a donor of genetic information. During conjugation, a copy of the F factor is almost always transferred from the F^+ cell to the F^- recipient, converting the recipient to the F^+ state. The question remained as to exactly why such a low proportion of these matings (10^{-7}) also results in genetic recombination. Also, it was unclear what the transfer of the F factor had to do with the transfer and recombination of particular genes. The answers to these questions awaited further experimentation.

As you soon shall see, the F factor is in reality an autonomous genetic unit referred to as a *plasmid*. However, in covering the history of its discovery, in this chapter we will continue to refer to it as a "factor."

Hfr Bacteria and Chromosome Mapping

Subsequent discoveries not only clarified how genetic recombination occurs but also defined a mechanism by which the *E. coli* chromosome could be mapped. Let's address chromosome mapping first.

In 1950, Cavalli-Sforza treated an F^+ strain of *E. coli* K12 with nitrogen mustard, a potent chemical known to induce mutations. From these treated cells, he recovered a genetically altered strain of donor bacteria that underwent recombination at a rate of $1/10^4$ (or 10^{-4}), 1000 times more frequently than the original F^+ strains. In 1953, William Hayes isolated another strain that demonstrated a similarly elevated frequency of recombination. Both strains were designated **Hfr,** for **high-frequency recombination.** Hfr cells constitute a special class of F^+ cells.

In addition to the higher frequency of recombination, another important difference was noted between Hfr strains and the original F^+ strains. If a donor cell is from an Hfr strain, recipient cells, though sometimes displaying genetic recombination, *never become Hfr;* thus they remain F^-. In comparison, then,

$$F^+ \times F^- \rightarrow \text{recipient becomes } F^+ \text{ (low rate of recombination)}$$

$$\text{Hfr} \times F^- \rightarrow \text{recipient remains } F^- \text{ (high rate of recombination)}$$

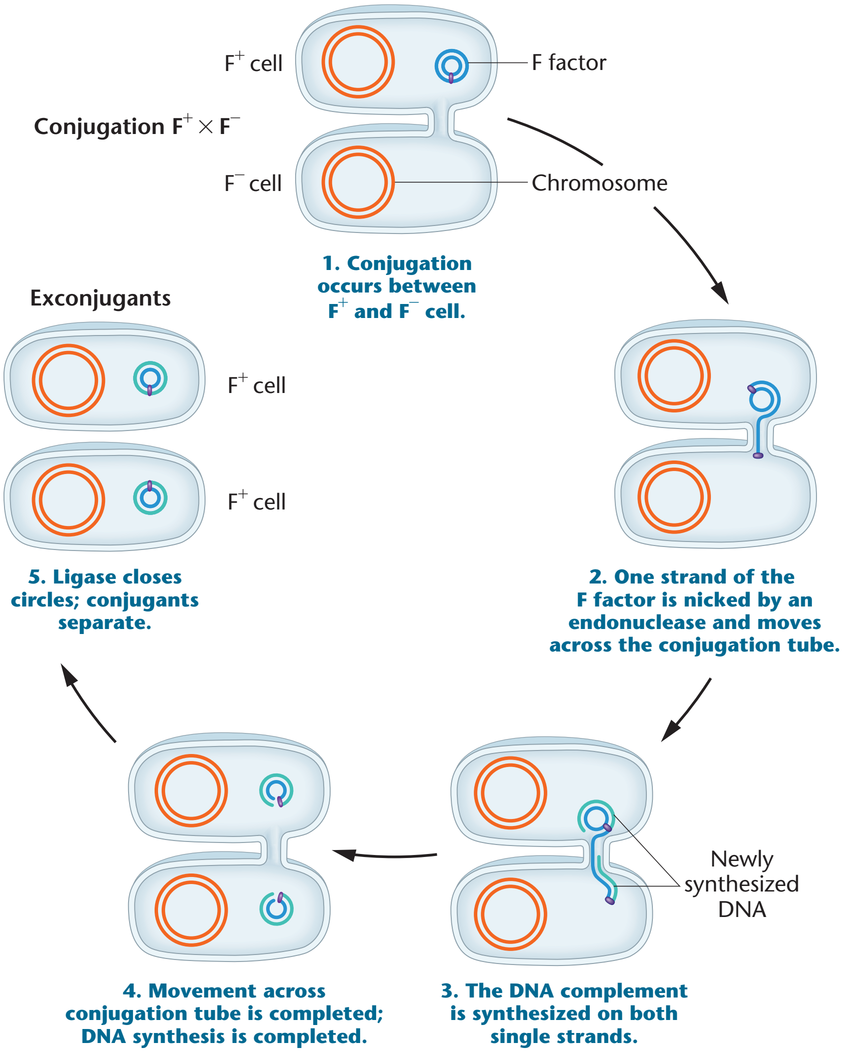

F+ cell
F factor
Conjugation F+ × F−
F− cell
Chromosome
1. Conjugation occurs between F+ and F− cell.
Exconjugants
F+ cell
F+ cell
5. Ligase closes circles; conjugants separate.
2. One strand of the F factor is nicked by an endonuclease and moves across the conjugation tube.
Newly synthesized DNA
4. Movement across conjugation tube is completed; DNA synthesis is completed.
3. The DNA complement is synthesized on both single strands.

FIGURE 6.5 An $F^+ \times F^-$ mating, demonstrating how the recipient F^- cell is converted to F^+. During conjugation, the DNA of the F factor is replicated, with one new copy entering the recipient cell, converting it to F^+.The bars drawn on the F factors indicate their clockwise rotation during replication. Newly replicated DNA is depicted by a lighter shade of blue as the F factor is transferred.

Perhaps the most significant characteristic of Hfr strains is the *specific nature of recombination*. In a given Hfr strain, certain genes are more frequently recombined than others, and some do not recombine at all. This *nonrandom* pattern of gene transfer was shown to vary among Hfr strains. Although these results were puzzling, Hayes interpreted them to mean that some physiological alteration of the F factor had occurred to produce Hfr strains of *E. coli*.

In the mid-1950s, experimentation by Ellie Wollman and François Jacob explained the differences between Hfr cells and F^+ cells and showed how Hfr strains would allow genetic mapping of the *E. coli* chromosome. In Wollman and Jacob's experiments, Hfr and antibiotic-resistant F^- strains with suitable marker genes were mixed, and recombination of these genes was assayed at different times. Specifically, a culture containing a mixture of an Hfr and an F^- strain was incubated, and samples were removed at intervals and placed in a blender. The shear forces created in the blender separated conjugating bacteria so that the transfer of the chromosome was terminated. Then the sampled cells were grown on medium containing the antibiotic, so that only recipient cells would be recovered. These cells were subsequently tested for the transfer of specific genes.

This process, called the **interrupted mating technique,** demonstrated that, depending on the specific Hfr strain, certain genes are transferred and recombined sooner than others. The graph in **Figure 6.6** illustrates this point. During the first 8 minutes after the two strains

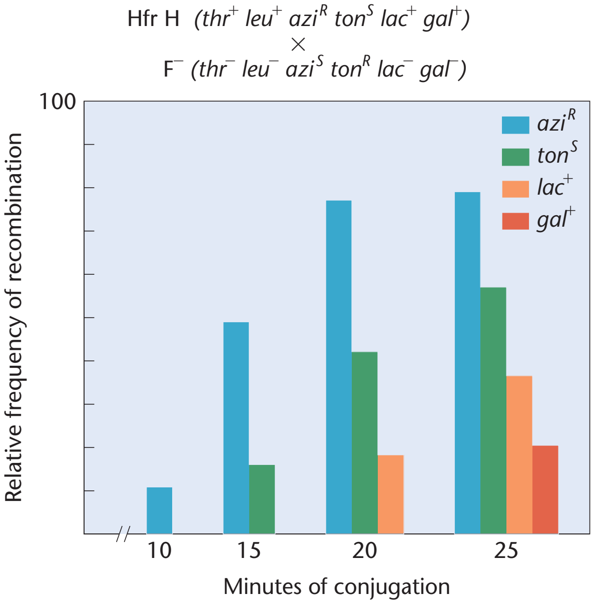


FIGURE 6.6 The progressive transfer during conjugation of various genes from a specific Hfr strain of *E. coli* to an F^- strain. Certain genes (*azi* and *ton*) transfer sooner than others and recombine more frequently. Others (*lac* and *gal*) transfer later, and recombinants are found at a lower frequency. Still others (*thr* and *leu*) are always transferred and were used in the initial screen for recombinants but are not shown here.

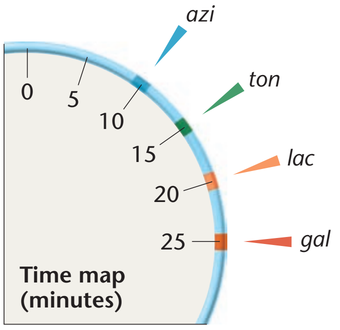


FIGURE 6.7 A time map of the genes studied in the experiment depicted in Figure 6.6.

were mixed, no genetic recombination was detected. At about 10 minutes, recombination of the azi^R gene could be detected, but no transfer of the ton^S, lac^+, or gal^+ genes was noted. By 15 minutes, 50 percent of the recombinants were azi^R and 15 percent were also ton^S; but none was lac^+ or gal^+. Within 20 minutes, the lac^+ gene was found among the recombinants; and within 25 minutes, gal^+ was also beginning to be transferred. Wollman and Jacob had demonstrated an *ordered transfer of genes* that correlated with the length of time conjugation proceeded.

It appeared that the chromosome of the Hfr bacterium was transferred linearly, so that the gene order and distance between genes, as measured in minutes, could be predicted from experiments such as Wollman and Jacob's (**Figure 6.7**). This information, sometimes referred to as **time mapping,** served as the basis for the first genetic map of the *E. coli* chromosome. Minutes in bacterial mapping provide a measure similar to map units in eukaryotes.

Wollman and Jacob repeated the same type of experiment with other Hfr strains, obtaining similar results but with one important difference. Although genes were always transferred linearly with time, as in their original experiment, the order in which genes entered the recipient seemed to vary from Hfr strain to Hfr strain [**Figure 6.8(a)**]. Nevertheless, when the researchers reexamined the entry rate of genes, and thus the different genetic maps for each strain, a distinct pattern emerged. The major difference between each strain was simply the point of the origin (*O*)—the first part of the donor chromosome to enter the recipient—and the direction in which entry proceeded from that point [**Figure 6.8(b)**].

To explain these results, Wollman and Jacob postulated that the *E. coli* chromosome is circular (a closed circle, with no free ends). If the point of origin (*O*) varies from strain to strain, a different sequence of genes will be transferred in each case. But what determines *O*? They proposed that, in various Hfr strains, the F factor integrates into the chromosome at different points, and its position determines the *O* site. One such case of integration is shown in step 1 of **Figure 6.9**. During conjugation between an Hfr and an F^- cell, the position of the F factor determines the initial point of transfer (steps 2 and 3). Those genes adjacent to *O* are transferred first, and the F factor becomes the last part that can be transferred (step 4). However, conjugation rarely, if ever, lasts long enough to allow the entire chromosome to pass across the conjugation tube (step 5). *This proposal explains why recipient cells, when mated with Hfr cells, remain* F^-.

Figure 6.9 also depicts the way in which the two strands making up a donor's DNA molecule behave during transfer, allowing for the entry of one strand of DNA into the recipient (step 3). Following its replication in the recipient, the entering DNA has the potential to recombine with the region homologous to it on the host chromosome. The DNA strand that remains in the donor also undergoes replication.

Use of the interrupted mating technique with different Hfr strains allowed researchers to map the entire *E. coli* chromosome. Mapped in time units, strain K12 (or

(a)

Hfr strain	(earliest)			Order of transfer				(latest)
H	*thr* –	*leu* –	*azi* –	*ton* –	*pro* –	*lac* –	*gal* –	*thi*
1	*leu* –	*thr* –	*thi* –	*gal* –	*lac* –	*pro* –	*ton* –	*azi*
2	*pro* –	*ton* –	*azi* –	*leu* –	*thr* –	*thi* –	*gal* –	*lac*
7	*ton* –	*azi* –	*leu* –	*thr* –	*thi* –	*gal* –	*lac* –	*pro*

(b)

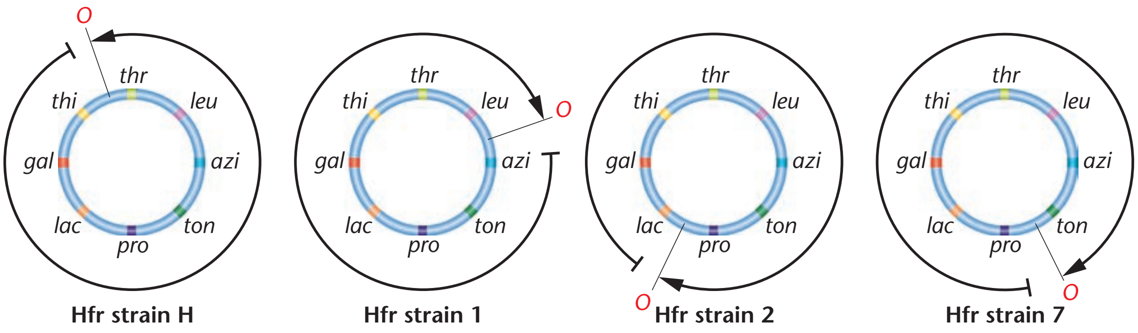


FIGURE 6.8 (a) The order of gene transfer in four Hfr strains, suggesting that the *E. coli* chromosome is circular. (b) The point where transfer originates (*O*) is identified in each strain. The origin is the point of integration of the F factor into the chromosome; the direction of transfer is determined by the orientation of the F factor as it integrates. The arrowheads indicate the points of initial transfer.

E. coli K12) was shown to be 100 minutes long. While modern genome analysis of the *E. coli* chromosome has now established the presence of just over 4000 protein-coding sequences, this original mapping procedure established the location of approximately 1000 genes.

NOW SOLVE THIS

6.1 When the interrupted mating technique was used with five different strains of Hfr bacteria, the following orders of gene entry and recombination were observed. On the basis of these data, draw a map of the bacterial chromosome. Do the data support the concept of circularity?

Hfr Strain			Order		
1	*T*	*C*	*H*	*R*	*O*
2	*H*	*R*	*O*	*M*	*B*
3	*M*	*O*	*R*	*H*	*C*
4	*M*	*B*	*A*	*K*	*T*
5	*C*	*T*	*K*	*A*	*B*

■ **HINT:** *This problem involves an understanding of how the bacterial chromosome is transferred during conjugation, leading to recombination and providing data for mapping. The key to its solution is to understand that chromosome transfer is strain-specific and depends on where in the chromosome, and in which orientation, the F factor has integrated.*

Recombination in $F^+ \times F^-$ Matings: A Reexamination

The preceding model helped geneticists better understand how genetic recombination occurs during the $F^+ \times F^-$ matings. Recall that recombination occurs much less frequently in them than in Hfr × F^- matings and that random gene transfer is involved. The current belief is that when F^+ and F^- cells are mixed, conjugation occurs readily, and each F^- cell involved in conjugation with an F^+ cell receives a copy of the F factor, *but no genetic recombination occurs*. However, at an extremely low frequency in a population of F^+ cells, the F factor integrates spontaneously into a random point in the bacterial chromosome, converting that F^+ cell to the Hfr state as shown in Figure 6.9. Therefore, in $F^+ \times F^-$ matings, the extremely low frequency of genetic recombination (10^{-7}) is attributed to the rare, newly formed Hfr cells, which then undergo conjugation with F^- cells. Because the point of integration of the F factor is random, the genes transferred by any newly formed Hfr donor *will also appear to be random within the larger* F^+/F^- *population*. The recipient bacterium will appear as a recombinant but will, in fact, remain F^-. If it subsequently undergoes conjugation with an F^+ cell, it will be converted to F^+.

The F′ State and Merozygotes

In 1959, during experiments with Hfr strains of *E. coli,* Edward Adelberg discovered that the F factor could lose its

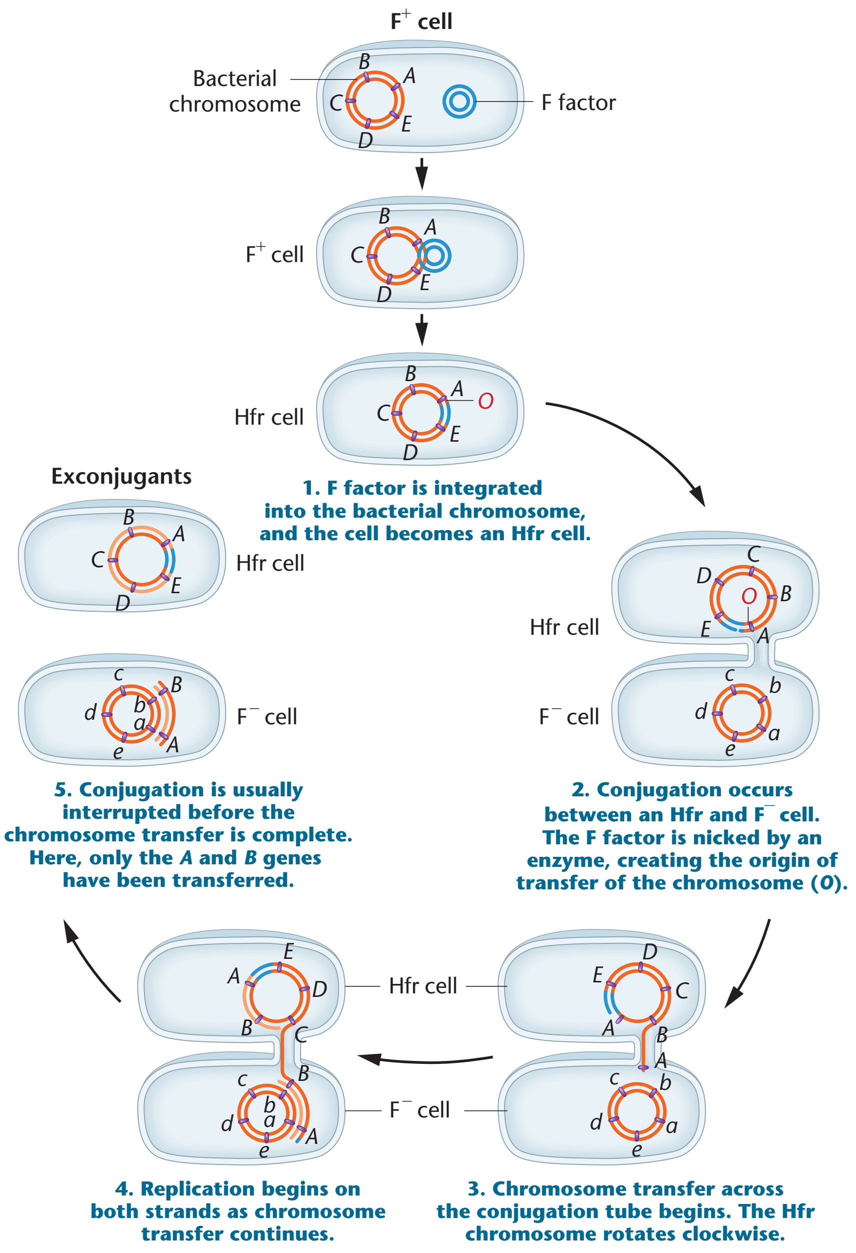
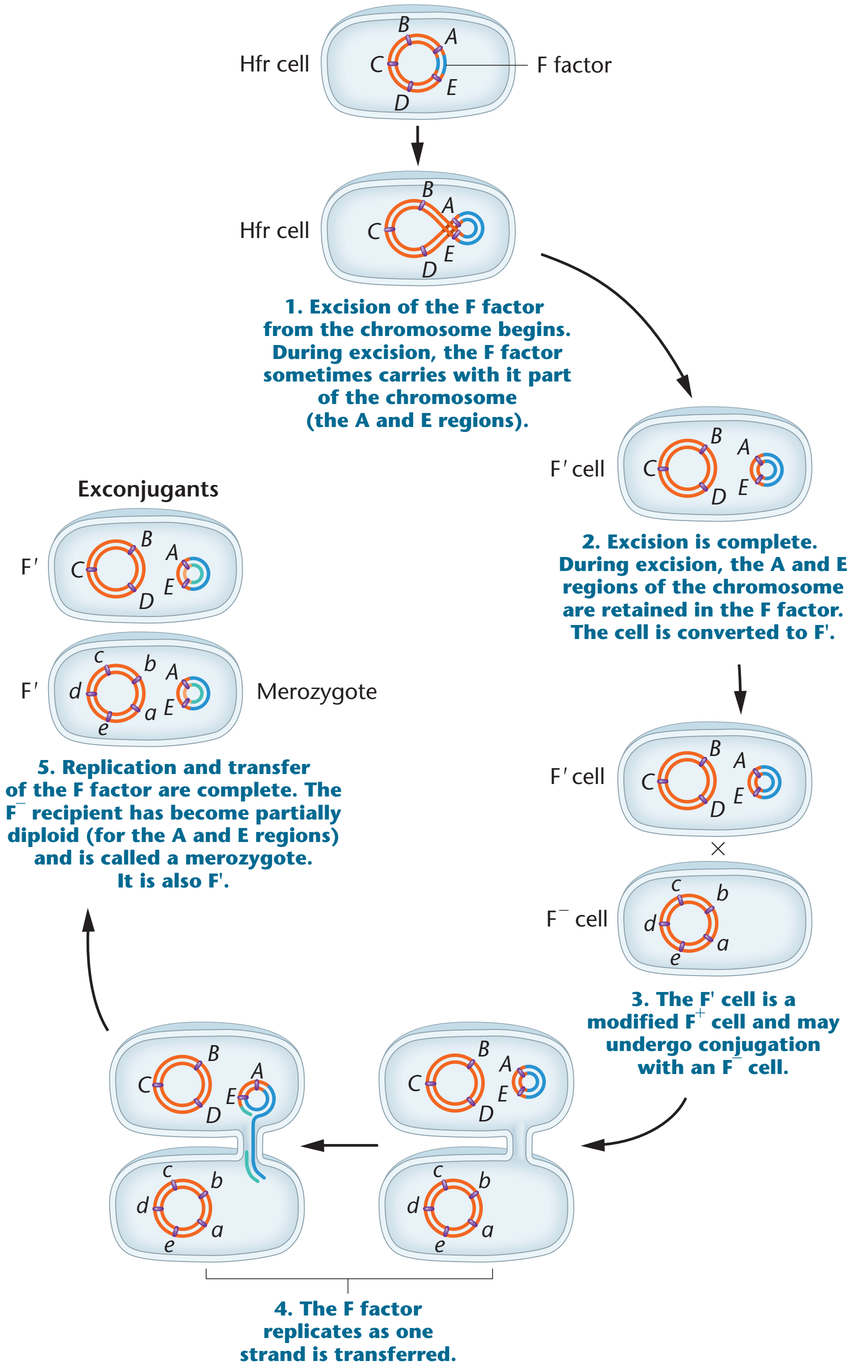

F+ cell
Bacterial chromosome
F factor
B
A
C
E
D
F+ cell
Hfr cell
O
1. F factor is integrated into the bacterial chromosome, and the cell becomes an Hfr cell.
Exconjugants
Hfr cell
F− cell
Hfr cell
F− cell
c
b
d
a
e
5. Conjugation is usually interrupted before the chromosome transfer is complete. Here, only the A and B genes have been transferred.
2. Conjugation occurs between an Hfr and F− cell. The F factor is nicked by an enzyme, creating the origin of transfer of the chromosome (O).
Hfr cell
F− cell
4. Replication begins on both strands as chromosome transfer continues.
3. Chromosome transfer across the conjugation tube begins. The Hfr chromosome rotates clockwise.

FIGURE 6.9 Conversion of F$^+$ to an Hfr state occurs by integration of the F factor into the bacterial chromosome. The point of integration determines the origin (*O*) of transfer. During conjugation, an enzyme nicks the F factor, now integrated into the host chromosome, initiating the transfer of the chromosome at that point. Conjugation is usually interrupted prior to complete transfer. Here, only the *A* and *B* genes are transferred to the F$^-$ cell; they may recombine with the host chromosome. Newly replicated DNA of the chromosome is depicted by a lighter shade of orange.

integrated status, causing the cell to revert to the F$^+$ state (**Figure 6.10**, step 1). When this occurs, the F factor frequently carries several adjacent bacterial genes along with it (step 2). Adelberg designated this condition F′ to distinguish it from F$^+$ and Hfr. F′, like Hfr, is thus another special case of F$^+$.

The presence of bacterial genes within a cytoplasmic F factor creates an interesting situation. An F′ bacterium behaves like an F$^+$ cell by initiating conjugation with F$^-$ cells (Figure 6.10, step 3). When this occurs, the F factor, containing chromosomal genes, is transferred to the F$^-$ cell (step 4). As a result, whatever chromosomal genes are part of the



1. Excision of the F factor from the chromosome begins. During excision, the F factor sometimes carries with it part of the chromosome (the A and E regions).

2. Excision is complete. During excision, the A and E regions of the chromosome are retained in the F factor. The cell is converted to F'.

3. The F' cell is a modified F^+ cell and may undergo conjugation with an F^- cell.

4. The F factor replicates as one strand is transferred.

5. Replication and transfer of the F factor are complete. The F^- recipient has become partially diploid (for the A and E regions) and is called a merozygote. It is also F'.

FIGURE 6.10 Conversion of an Hfr bacterium to F′ and its subsequent mating with an F^- cell. The conversion occurs when the F factor loses its integrated status. During excision from the chromosome, the F factor may carry with it one or more chromosomal genes (in this case, *A* and *E*). Following conjugation, the recipient cell becomes partially diploid and is called a merozygote. It also behaves as an F^+ donor cell.

F factor are now present as duplicates in the recipient cell (step 5) because the recipient still has a complete chromosome. This creates a partially diploid cell called a **merozygote.** Pure cultures of F′ merozygotes can be established. They have been extremely useful in the study of genetic regulation in bacteria, as we will discuss later in the text (see Chapter 16).

6.3 The F Factor Is an Example of a Plasmid

The preceding sections introduced the extrachromosomal heredity unit called the F factor that bacteria require for conjugation. When it exists autonomously in the bacterial

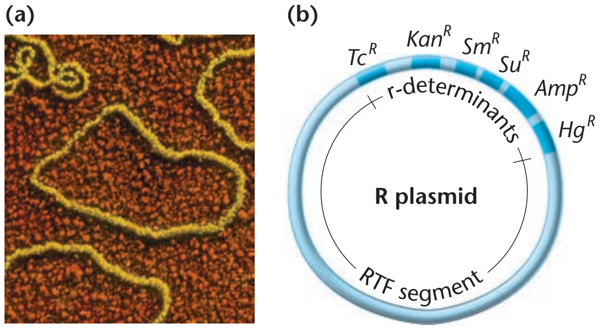


FIGURE 6.11 (a) Electron micrograph of plasmids isolated from *E. coli.* (b) An R plasmid containing a resistance transfer factor (RTF) and multiple r-determinants (Tc, tetracycline; Kan, kanamycin; Sm, streptomycin; Su, sulfonamide; Amp, ampicillin; and Hg, mercury).

cytoplasm, it is composed of a double-stranded closed circle of DNA. These characteristics place the F factor in the more general category of genetic structures called **plasmids** [Figure 6.11(a)]. Plasmids often exist in multiple copies in the cytoplasm; each may contain one or more genes and often quite a few. Their replication depends on the same enzymes that replicate the chromosome of the host cell, and they are distributed to daughter cells along with the host chromosome during cell division. Many plasmids are confined to the cytoplasm of the bacterial cell. Others, such as the F factor, can integrate into the host chromosome. Those plasmids that can exist autonomously or can integrate into the chromosome are further designated as **episomes**.

Plasmids can be classified according to the genetic information specified by their DNA. The F factor plasmid confers fertility and contains genes essential for sex pilus formation, on which conjugation and subsequent genetic recombination depend. Other examples of plasmids include the R and the Col plasmids.

Most **R plasmids** consist of two components: the **resistance transfer factor (RTF)** and one or more **r-determinants** [Figure 6.11(b)]. The RTF encodes genetic information essential to transferring the plasmid between bacteria, and the r-determinants are genes conferring resistance to antibiotics or heavy metals such as mercury. While RTFs are quite similar in a variety of plasmids from different bacterial species, there is wide variation in r-determinants, each of which is specific for resistance to one class of antibiotic. Determinants with resistance to tetracycline, streptomycin, ampicillin, sulfonamide, kanamycin, or chloramphenicol are the most frequently encountered. Sometimes plasmids contain many r-determinants, conferring resistance to several antibiotics [Figure 6.11(b)]. Bacteria bearing such plasmids are of great medical significance, not only because of their multiple resistance but also because of the ease with which the plasmids may be transferred to other pathogenic bacteria, rendering those bacteria resistant to a wide range of antibiotics.

The first known case of such a plasmid occurred in Japan in the 1950s in the bacterium *Shigella*, which causes dysentery. In hospitals, bacteria were isolated that were resistant to as many as five of the above antibiotics. Obviously, this phenomenon represents a major health threat. Fortunately, a bacterial cell sometimes contains r-determinant plasmids but no RTF. Although such a cell is resistant, it cannot transfer the genetic information for resistance to recipient cells. The most commonly studied plasmids, however, contain the RTF as well as one or more r-determinants.

The **Col plasmid,** ColE1 (derived from *E. coli*), is clearly distinct from R plasmids. It encodes one or more proteins that are highly toxic to bacterial strains that do not harbor the same plasmid. These proteins, called **colicins,** can kill neighboring bacteria, and bacteria that carry the plasmid are said to be *colicinogenic*. Present in 10 to 20 copies per cell, the Col plasmid also contains a gene encoding an immunity protein that protects the host cell from the toxin. Unlike an R plasmid, the Col plasmid is not usually transmissible to other cells.

Interest in plasmids has increased dramatically because of their role in recombinant DNA research. As we will see (in Chapter 20), specific genes from any source can be inserted into a plasmid, which may then be inserted into a bacterial cell. As the altered cell replicates its DNA and undergoes division, the foreign gene is also replicated, thus being cloned.

6.4 Transformation Is a Second Process Leading to Genetic Recombination in Bacteria

Transformation provides another mechanism for recombining genetic information in some bacteria. Small pieces of extracellular (exogenous) DNA are taken up by a living bacterium, potentially leading to a stable genetic change in the recipient cell. We discuss transformation in this chapter because in those bacterial species in which it occurs, the process can be used to map bacterial genes, though in a more limited way than conjugation. As we will see later in the text (see Chapter 10), the process of transformation was also instrumental in proving that DNA is the genetic material. Further, in recombinant DNA studies (Chapter 20), transformation, albeit an artificial version often enhanced by *electroporation* (electric current), is instrumental in gene cloning.

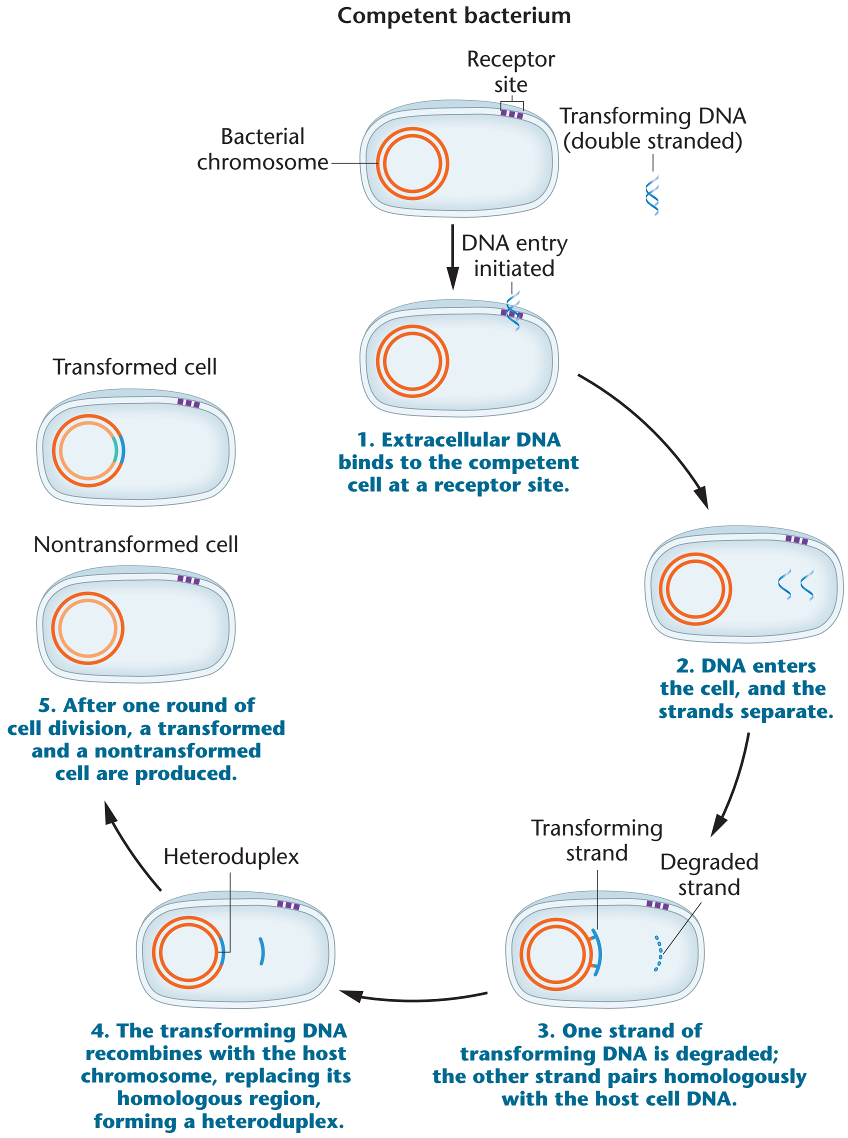

Competent bacterium
Receptor site
Bacterial chromosome
Transforming DNA (double stranded)
DNA entry initiated
1. Extracellular DNA binds to the competent cell at a receptor site.
2. DNA enters the cell, and the strands separate.
Transforming strand
Degraded strand
3. One strand of transforming DNA is degraded; the other strand pairs homologously with the host cell DNA.
Heteroduplex
4. The transforming DNA recombines with the host chromosome, replacing its homologous region, forming a heteroduplex.
Transformed cell
Nontransformed cell
5. After one round of cell division, a transformed and a nontransformed cell are produced.

FIGURE 6.12 Proposed steps for transformation of a bacterial cell by exogenous DNA. Only one of the two strands of the entering DNA is involved in the transformation event, which is completed following cell division.

The Transformation Process

The process of transformation (Figure 6.12) consists of numerous steps that achieve two basic outcomes: (1) entry of foreign DNA into a recipient cell; and (2) recombination between the foreign DNA and its homologous region in the recipient chromosome. While completion of both outcomes is required for genetic recombination, the first step of transformation can occur without the second step, resulting in the addition of foreign DNA to the bacterial cytoplasm but not to its chromosome.

In a population of bacterial cells, only those in a particular physiological state of **competence** take up DNA. Studies have shown that various kinds of bacteria readily undergo transformation naturally (e.g., *Haemophilus influenzae, Bacillus subtilis, Shigella paradysenteriae, Streptococcus pneumoniae,* and *E. coli*). Others can be induced in the laboratory to become competent. Entry of DNA is thought to occur at a limited number of receptor sites on the surface of a competent bacterial cell (Figure 6.12, step 1). Passage into the cell is thought to be an active process that requires energy and specific transport molecules. This model is supported by the fact that substances that inhibit energy production or protein synthesis also inhibit transformation.

Soon after entry, one of the two strands of the double helix is digested by nucleases, leaving only a single strand to participate in transformation (Figure 6.12, steps 2 and 3). The surviving DNA strand aligns with the complementary region of the bacterial chromosome. In a process involving several enzymes, this segment of DNA replaces its counterpart in the chromosome (step 4), which is excised and degraded.

For recombination to be detected, the transforming DNA must be derived from a different strain of bacteria that bears some distinguishing genetic variation, such as a mutation. Once this is integrated into the chromosome, the recombinant region contains one host strand (present originally) and one mutant strand. Because these strands are from different sources, the region is referred to as a **heteroduplex,** which usually contains some mismatch of base sequence. This mismatch activates a repair process (see Chapter 15). Following repair and one round of DNA replication, one chromosome is restored to its original DNA sequence, identical to that of the original recipient cell, and the other contains the properly aligned mutant gene. Following cell division, one nontransformed cell (nonmutant) and one transformed cell (mutant) are produced (step 5).

Transformation and Linked Genes

In early transformation studies, the most effective exogenous DNA was a size containing 10,000–20,000 nucleotide pairs, a length sufficient to encode several genes.* Genes adjacent

* Today, we know that a 2000 nucleotide pair length of DNA is highly effective in gene cloning experiments.

NOW SOLVE THIS

6.2 In a transformation experiment involving a recipient bacterial strain of genotype a^-b^-, the following results were obtained. What can you conclude about the location of the *a* and *b* genes relative to each other?

	Transformants (%)		
Transforming DNA	a^+b^-	a^-b^+	a^+b^+
a^+b^+	3.1	1.2	0.04
a^+b^- and a^-b^+	2.4	1.4	0.03

■ **HINT:** *This problem involves an understanding of how transformation can be used to determine if bacterial genes are closely "linked." You are asked to predict the location of two genes relative to one another. The key to its solution is to understand that cotransformation (of two genes) occurs according to the laws of probability. Two "unlinked" genes are transformed only as a result of two independent events. In such a case, the probability of that occurrence is equal to the product of the individual probabilities.*

to or very close to one another on the bacterial chromosome can be carried on a single segment of this size. Consequently, a single transfer event can result in the **cotransformation** of several genes simultaneously. Genes that are close enough to each other to be cotransformed are *linked*. In contrast to *linkage groups* in eukaryotes, which consist of all genes on a single chromosome, note that here *linkage* refers to the proximity of genes that permits cotransformation (i.e., the genes are next to, or close to, one another).

If two genes are not linked, simultaneous transformation occurs only as a result of two independent events involving two distinct segments of DNA. As with double crossovers in eukaryotes, the probability of two independent events occurring simultaneously is equal to the product of the individual probabilities. Thus, the frequency of two unlinked genes being transformed simultaneously is much lower than if they are linked. Under certain conditions, relative distances between linked genes can be determined from transformation data in a manner analogous to chromosome mapping in eukaryotes, though somewhat more complex.

6.5 Bacteriophages Are Bacterial Viruses

Bacteriophages, or **phages** as they are commonly known, are viruses that have bacteria as their hosts. The reproduction of phages can lead to still another mode of bacterial genetic recombination, called transduction. To understand this process, we first must consider the genetics of bacteriophages, which themselves can undergo recombination.

A great deal of genetic research has been done using bacteriophages as a model system. In this section, we will first examine the structure and life cycle of one type of bacteriophage. We then discuss how these phages are studied during their infection of bacteria. Finally, we contrast two possible modes of behavior once initial phage infection occurs. This information is background for our discussion of *transduction* and *bacteriophage recombination*.

Phage T4: Structure and Life Cycle

Bacteriophage T4, which has *E. coli* as its host, is one of a group of related bacterial viruses referred to as T-even phages. It exhibits the intricate structure shown in **Figure 6.13**. Its genetic material, DNA, is contained within an icosahedral (referring to a polyhedron with 20 faces) protein coat, making up the head of the virus. The DNA is sufficient in quantity to encode more than 150 average-sized genes. The head is connected to a complex tail structure consisting of a collar, an outer contractile sheath that surrounds an inner spike-like tube, which sits atop a base plate from which tail fibers protrude. The base plate is an extremely complex structure, consisting of 15 different proteins, most present in multiple copies. The base plate coordinates the host cell recognition and is involved in providing the signal whereby the outer sheath contracts, propelling the inner tube across the cell membrane of the host cell.

The life cycle of phage T4 (**Figure 6.14**) is initiated when the virus binds to the bacterial host cell. Then, during contraction of the outer sheath, the DNA in the head is extruded, and it moves across the cell membrane into the bacterial cytoplasm. Within minutes, all bacterial DNA, RNA, and protein synthesis is inhibited, and synthesis of viral molecules begins. At the same time, degradation of the host DNA is initiated.

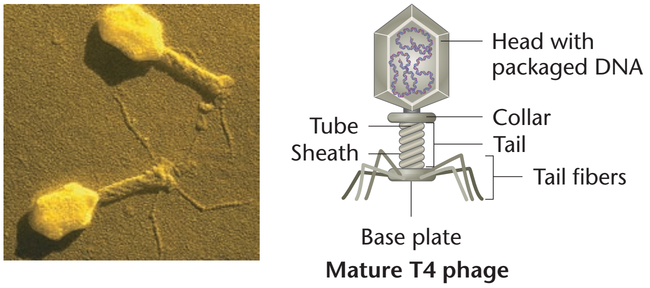


FIGURE 6.13 The structure of bacteriophage T4, which includes an icosahedral head filled with DNA; a tail consisting of a collar, tube, and sheath; and a base plate with tail fibers. During assembly, the tail components are added to the head and then tail fibers are added.

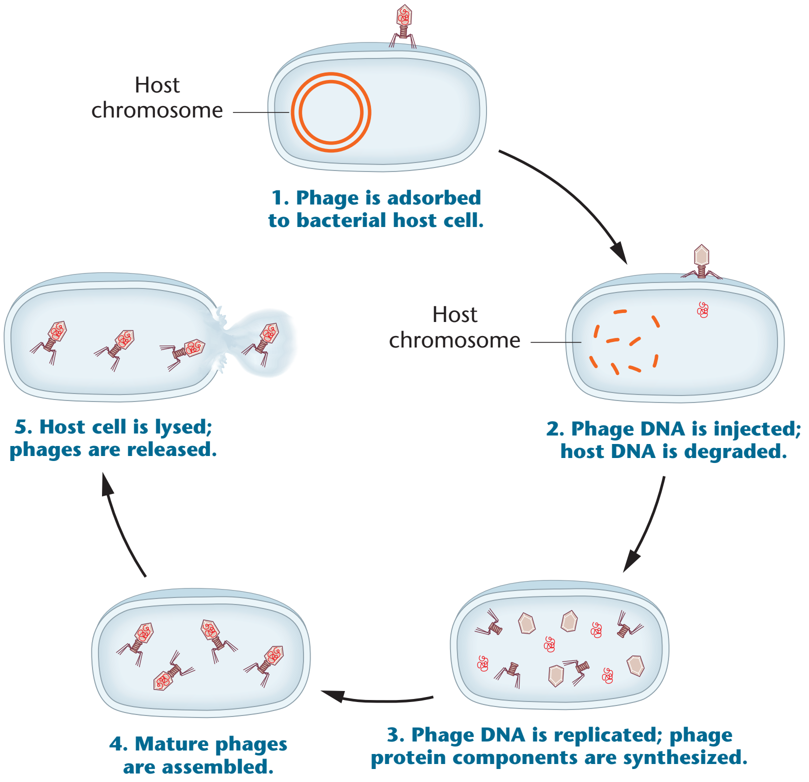

Host chromosome
1. Phage is adsorbed to bacterial host cell.
Host chromosome
2. Phage DNA is injected; host DNA is degraded.
3. Phage DNA is replicated; phage protein components are synthesized.
4. Mature phages are assembled.
5. Host cell is lysed; phages are released.

FIGURE 6.14 Life cycle of bacteriophage T4.

A period of intensive viral gene activity characterizes infection. Initially, phage DNA replication occurs, leading to a pool of viral DNA molecules. Then, the components of the head, tail, and tail fibers are synthesized. The assembly of mature viruses is a complex process that has been well studied by William Wood, Robert Edgar, and others. Three sequential pathways take part: (1) DNA packaging as the viral heads are assembled, (2) tail assembly, and (3) tail-fiber assembly. Once DNA is packaged into the head, that structure combines with the tail components, to which tail fibers are added. Total construction is a combination of self-assembly and enzyme-directed processes.

When approximately 200 new viruses are constructed, the bacterial cell is ruptured by the action of lysozyme (a phage gene product), and the mature phages are released from the host cell. This step during infection is referred to as **lysis,** and it completes what is referred to as the **lytic cycle.** The 200 new phages infect other available bacterial cells, and the process repeats itself over and over again.

The Plaque Assay

Bacteriophages and other viruses have played a critical role in our understanding of molecular genetics. During infection of bacteria, enormous quantities of bacteriophages may be obtained for investigation. Often, more than 10^{10} viruses are produced per milliliter of culture medium. Many genetic studies have relied on our ability to determine the number of phages produced following infection under specific culture conditions. The **plaque assay,** routinely used for such determinations, is invaluable in quantitative analysis during mutational and recombinational studies of bacteriophages.

This assay is illustrated in **Figure 6.15**, where actual plaque morphology is also shown. A serial dilution of the original virally infected bacterial culture is performed. Then, a 0.1-mL sample (an aliquot, meaning a fractional portion) from a dilution is added to a small volume of melted nutrient agar (about 3 mL) into which a few drops of a healthy bacterial culture have been added. The solution is then poured evenly over a base of solid nutrient agar in a petri dish and allowed to solidify before incubation. A clear area called a **plaque** occurs wherever a single virus initially infected one bacterium in the culture (the lawn) that has grown up during incubation. The plaque represents clones of the single infecting bacteriophage, created as reproduction cycles are repeated. If the dilution factor is too low, the plaques will be plentiful, and they may fuse, lysing the entire lawn of bacteria. This has occurred in the 10^{-3} dilution in Figure 6.15. However, if the dilution factor is increased appropriately, plaques can be counted, and the density of viruses in the initial culture can be estimated:

$$\text{initial phage density} = (\text{plaque number/mL}) \times (\text{dilution factor})$$

Figure 6.15 shows that 23 phage plaques were derived from the 0.1-mL aliquot of the 10^{-5} dilution. Therefore, we estimate a density of 230 phages/mL *at this dilution* (since the initial aliquot was 0.1 mL). The initial phage density in the undiluted sample, given that 23 plaques were observed from 0.1 mL of the 10^{-5} dilution, is calculated as

$$\text{initial phage density} = (230/\text{mL}) \times (10^5) = (230 \times 10^5)/\text{mL}$$

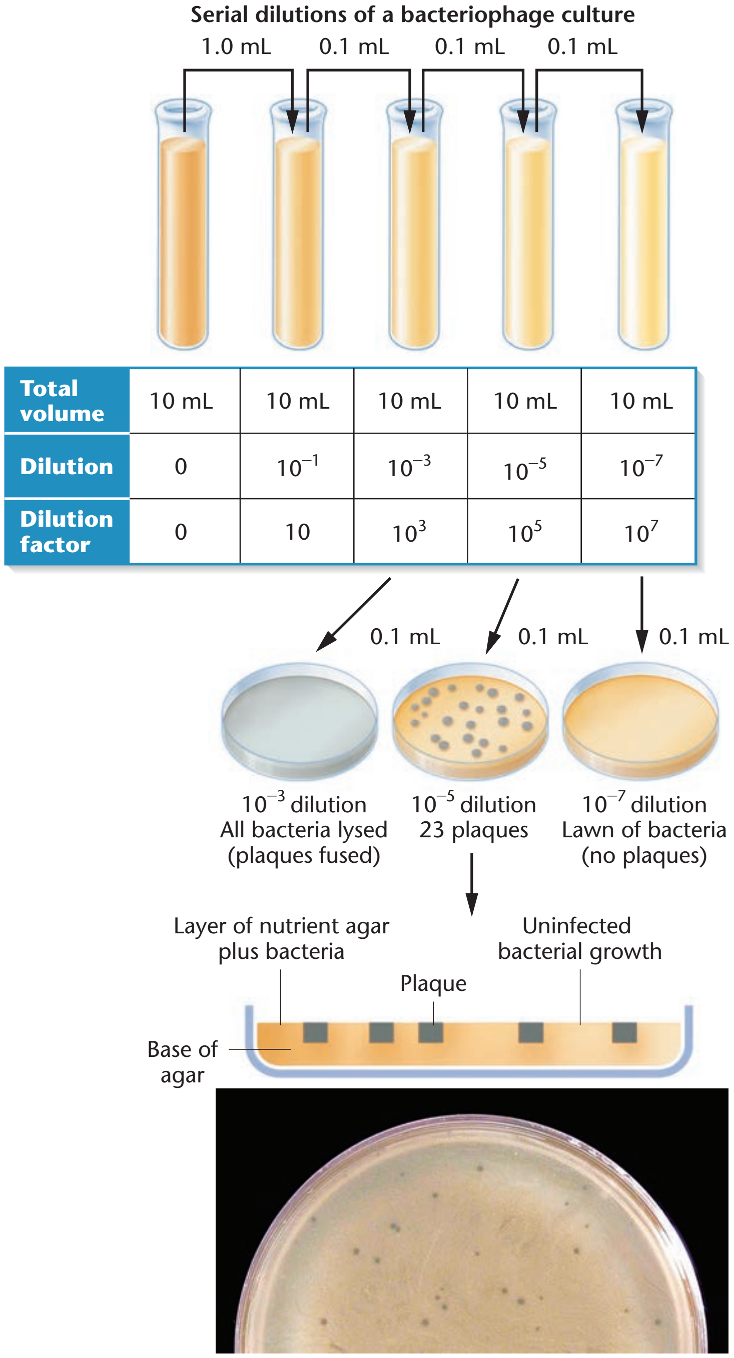
Because this figure is derived from the 10^{-5} dilution, we can also estimate that there would be only 0.23 phage/0.1 mL in the 10^{-7} dilution. Thus, if 0.1 mL from this tube were assayed, we would predict that no phage particles would be present. This prediction is borne out in Figure 6.15, where an intact lawn of bacteria lacking any plaques is depicted. The dilution factor is simply too great.

Use of the plaque assay has been invaluable in mutational and recombinational studies of bacteriophages. We will apply this technique more directly later in this chapter when we discuss Seymour Benzer's elegant genetic analysis of a single gene in phage T4.

Lysogeny

Infection of a bacterium by a virus does not always result in viral reproduction and lysis. As early as the 1920s, it was known that a virus can enter a bacterial cell and coexist with it. The precise molecular basis of this relationship is now well understood. Upon entry, the viral DNA is integrated into the bacterial chromosome instead of replicating in the bacterial cytoplasm; this integration characterizes the developmental stage referred to as **lysogeny.** Subsequently, each time the bacterial chromosome is replicated, the viral DNA is also replicated and passed to daughter bacterial cells following division. No new viruses are produced, and no lysis of the bacterial cell occurs. However, under certain stimuli, such as chemical or ultraviolet-light treatment, the viral DNA loses its integrated status and initiates replication, phage reproduction, and lysis of the bacterium.

Several terms are used in describing this relationship. The viral DNA integrated into the bacterial chromosome is called a **prophage.** Viruses that can either lyse the cell or behave as a prophage are called **temperate phages**. Those that can only lyse the cell are referred to as **virulent phages.** A bacterium harboring a prophage has been **lysogenized** and is said to be **lysogenic;** that is, it is capable of



Total volume	10 mL	10 mL	10 mL	10 mL	10 mL
Dilution	0	10^{-1}	10^{-3}	10^{-5}	10^{-7}
Dilution factor	0	10	10^3	10^5	10^7

FIGURE 6.15 A plaque assay for bacteriophage analysis. First, serial dilutions are made of a bacterial culture infected with bacteriophages. Then, three of the dilutions (10^{-3}, 10^{-5}, and 10^{-7}) are analyzed using the plaque assay technique. Each plaque represents the initial infection of one bacterial cell by one bacteriophage. In the 10^{-3} dilution, so many phages are present that all bacteria are lysed. In the 10^{-5} dilution, 23 plaques are produced. In the 10^{-7} dilution, the dilution factor is so great that no phages are present in the 0.1-mL sample, and thus no plaques form. From the 0.1-mL sample of the 10^{-5} dilution, the original bacteriophage density is calculated to be 23 × 10 × 10^5 phages/mL (230 × 10^5). The photograph shows phage T2 plaques on lawns of *E. coli.*

being lysed as a result of induced viral reproduction. The viral DNA (like the F factor discussed earlier) is classified as an *episome*, meaning a genetic molecule that can replicate either in the cytoplasm of a cell or as part of its chromosome.

6.6 Transduction Is Virus-Mediated Bacterial DNA Transfer

In 1952, Norton Zinder and Joshua Lederberg were investigating possible recombination in the bacterium *Salmonella typhimurium*. Although they recovered prototrophs from mixed cultures of two different auxotrophic strains, subsequent investigations showed that recombination was not due to the presence of an F factor and conjugation, as in *E. coli*. What they discovered was a process of bacterial recombination mediated by bacteriophages and now called **transduction**.

The Lederberg–Zinder Experiment

Lederberg and Zinder mixed the *Salmonella* auxotrophic strains LA-22 and LA-2 together, and when the mixture was plated on minimal medium, they recovered prototrophic cells. The LA-22 strain was unable to synthesize the amino acids phenylalanine and tryptophan (*phe*$^-$*trp*$^-$), and LA-2 could not synthesize the amino acids methionine and histidine (*met*$^-$*his*$^-$). Prototrophs (*phe*$^+$*trp*$^+$ *met*$^+$*his*$^+$) were recovered at a rate of about $1/10^5$ (or 10^{-5}) cells.

Although these observations at first suggested that the recombination was the type observed earlier in conjugative strains of *E. coli*, experiments using the Davis U-tube soon showed otherwise (**Figure 6.16**). The two auxotrophic strains were separated by a sintered glass filter, thus preventing contact between the strains while allowing them to grow in a common medium. Surprisingly, when samples were removed from both sides of the filter and plated independently on minimal medium, prototrophs *were* recovered, but only from the side of the tube containing LA-22 bacteria. Recall that if conjugation were responsible, the Davis U-tube should have *prevented* recombination altogether (see Figure 6.4).

Since LA-2 cells appeared to be the source of the new genetic information (*phe*$^+$ and *trp*$^+$), how that information crossed the filter from the LA-2 cells to the LA-22 cells, allowing recombination to occur, was a mystery. The unknown source was designated simply as a **filterable agent (FA)**.

Three observations were used to identify the FA:

1. The FA was produced by the LA-2 cells only when they were grown in association with LA-22 cells. If LA-2 cells were grown independently in a culture medium that was later added to LA-22 cells, recombination did not occur. Therefore, the LA-22 cells played some role in the production of FA by LA-2 cells but did so only when the two strains were sharing a common growth medium.
2. The addition of DNase, which enzymatically digests DNA, did not render the FA ineffective. Therefore, the FA is not exogenous DNA, ruling out transformation.
3. The FA could not pass across the filter of the Davis U-tube when the pore size was reduced below the size of bacteriophages.

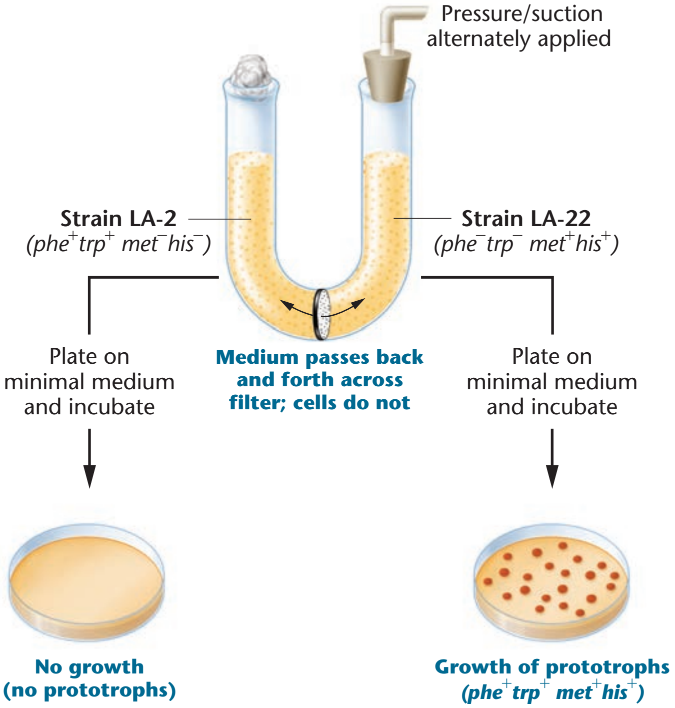

Pressure/suction alternately applied
Strain LA-2 (*phe*$^+$*trp*$^+$ *met*$^-$*his*$^-$)
Strain LA-22 (*phe*$^-$*trp*$^-$ *met*$^+$*his*$^+$)
Plate on minimal medium and incubate
Medium passes back and forth across filter; cells do not
Plate on minimal medium and incubate
No growth (no prototrophs)
Growth of prototrophs (*phe*$^+$*trp*$^+$ *met*$^+$*his*$^+$)

FIGURE 6.16 The Lederberg–Zinder experiment using *Salmonella*. After placing two auxotrophic strains on opposite sides of a Davis U-tube, Lederberg and Zinder recovered prototrophs from the side with the LA-22 strain but not from the side containing the LA-2 strain.

Aided by these observations and aware that temperate phages could lysogenize *Salmonella,* researchers proposed that the genetic recombination event was mediated by bacteriophage P22, present initially as a prophage in the chromosome of the LA-22 *Salmonella* cells. They hypothesized that P22 prophages sometimes enter the vegetative, or lytic, phase, reproduce, and are released by the LA-22 cells. Such P22 phages, being much smaller than a bacterium, then cross the filter of the U-tube and subsequently infect and lyse some of the LA-2 cells. In the process of lysis of LA-2, the P22 phages occasionally package a region of the LA-2 chromosome in their heads. If this region contains the *phe*$^+$ and *trp*$^+$ genes, and if the phages subsequently pass back across the filter and infect LA-22 cells, these newly lysogenized cells will behave as prototrophs. This process of transduction, whereby bacterial recombination is mediated by bacteriophage P22, is diagrammed in **Figure 6.17**.

Transduction and Mapping

Like transformation, transduction has been used in linkage and mapping studies of the bacterial chromosome. The fragment of bacterial DNA involved in a transduction event may be large enough to include several genes. As a result, two genes that are close to one another along the bacterial chromosome (i.e., are linked) can be transduced simultaneously, a process called **cotransduction.** If two genes are not close enough to one another along the chromosome to be included on a single DNA fragment, two independent transduction events must occur to carry them into a single cell. Since this occurs with a much lower probability than cotransduction, linkage can be determined by comparing the frequency of specific simultaneous recombinations.

By concentrating on two or three linked genes, transduction studies can also determine the precise order of these genes. The closer linked genes are to each other, the greater the frequency of cotransduction. Mapping studies can be done on three closely aligned genes, predicated on the same rationale that underlies other mapping techniques.

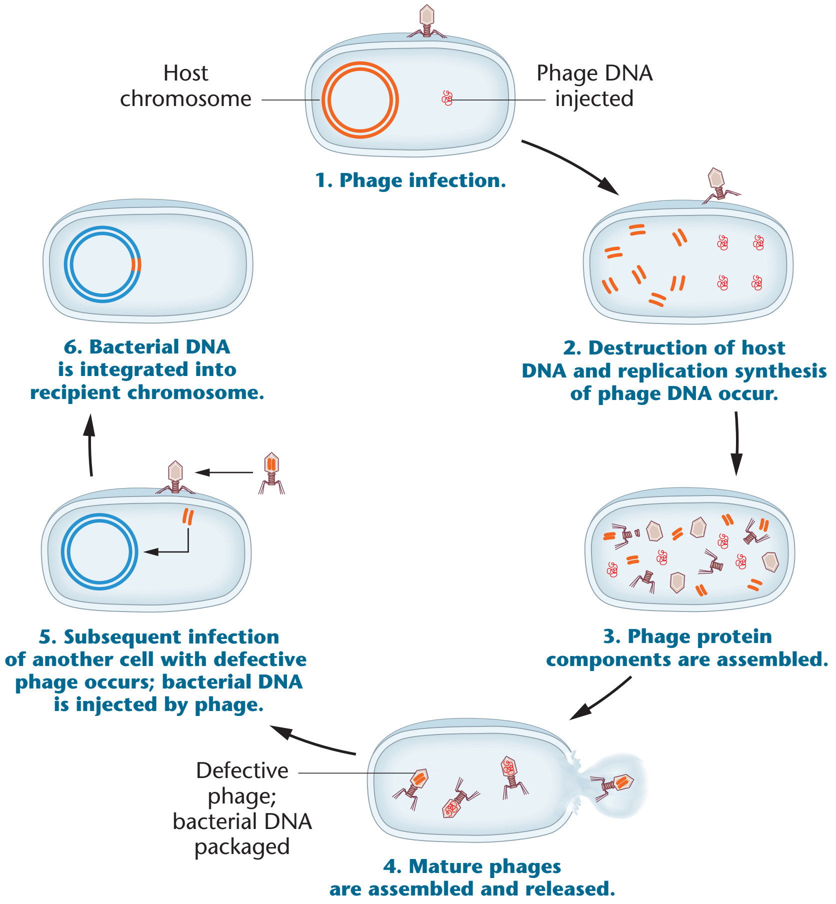

Host chromosome
Phage DNA injected
1. Phage infection.
2. Destruction of host DNA and replication synthesis of phage DNA occur.
3. Phage protein components are assembled.
4. Mature phages are assembled and released.
Defective phage; bacterial DNA packaged
5. Subsequent infection of another cell with defective phage occurs; bacterial DNA is injected by phage.
6. Bacterial DNA is integrated into recipient chromosome.

FIGURE 6.17 Generalized transduction.

6.7 Bacteriophages Undergo Intergenic Recombination

Around 1947, several research teams demonstrated that genetic recombination can be detected in bacteriophages. This led to the discovery that gene mapping can be performed in these viruses. Such studies relied on finding numerous phage mutations that could be visualized or assayed. As in bacteria and eukaryotes, these mutations allow genes to be identified and followed in mapping experiments. Before considering recombination and mapping in these bacterial viruses, we briefly introduce several of the mutations that were studied.

Bacteriophage Mutations

Phage mutations often affect the morphology of the plaques formed following lysis of bacterial cells. For example, in 1946, Alfred Hershey observed unusual T2 plaques on plates of *E. coli* strain B. Normal T2 plaques are small and have a clear center surrounded by a diffuse (nearly invisible) halo. In contrast, the unusual plaques were larger and possessed a distinctive outer perimeter (compare the lighter plaques in **Figure 6.18**). When the viruses were isolated from these plaques and replated on *E. coli* B cells, the resulting plaque appearance was identical. Thus, the plaque phenotype was an inherited trait resulting from the reproduction of mutant phages. Hershey named the mutant *rapid lysis* (*r*) because the plaques' larger size was thought to be due to a more rapid or more efficient life cycle of the phage. We now know that, in wild-type phages, reproduction is inhibited once a particular-sized plaque has been formed. The *r* mutant T2 phages overcome this inhibition, producing larger plaques.

Salvador Luria discovered another bacteriophage mutation, *host range* (*h*). This mutation extends the range of bacterial hosts that the phage can infect. Although wild-type T2 phages can infect *E. coli* B (a unique strain), they normally cannot attach or be adsorbed to the surface of *E. coli* B-2 (a different strain). The *h* mutation, however, confers the ability to adsorb to and subsequently infect *E. coli* B-2. When grown on a mixture of *E. coli* B and B-2, the *h* plaque has a center that appears much darker than that of the h^+ plaque (Figure 6.18).

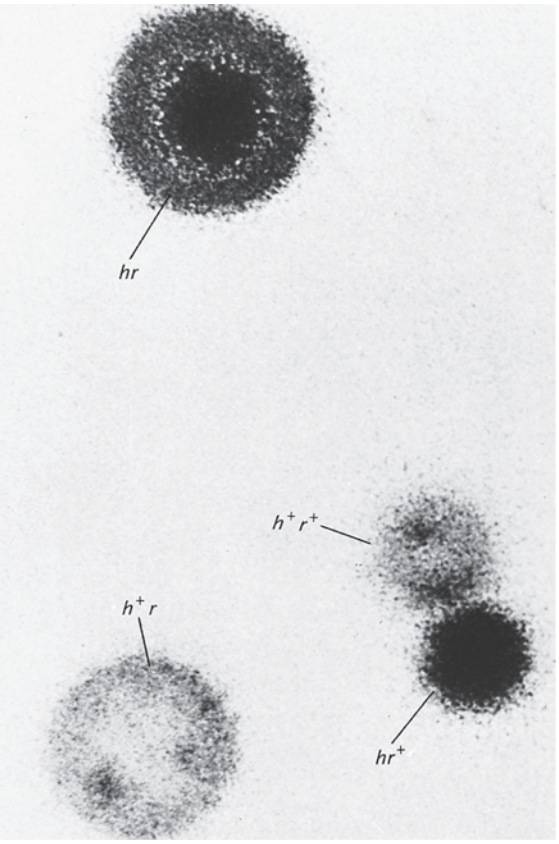


FIGURE 6.18 Plaque morphology phenotypes observed following simultaneous infection of *E. coli* by two strains of phage T2, h^+r and hr^+. In addition to the parental genotypes, recombinant plaques *hr* and h^+r^+ are shown.

Table 6.1 lists other types of mutations that have been isolated and studied in the T-even series of bacteriophages (e.g., T2, T4, T6). These mutations are important to the study of genetic phenomena in bacteriophages.

Mapping in Bacteriophages

Genetic recombination in bacteriophages was discovered during **mixed infection experiments,** in which two distinct mutant strains were allowed to *simultaneously* infect the same bacterial culture. These studies were designed so that the number of viral particles sufficiently exceeded the number of bacterial cells to ensure simultaneous infection of most cells by both viral strains. If two loci are involved, recombination is referred to as **intergenic**.

For example, in one study using the T2/*E. coli* system, the parental viruses were of either the h^+r (wild-type host range, rapid lysis) or the hr^+ (extended host range, normal lysis) genotype. If no recombination occurred, these two parental genotypes would be the only expected phage progeny. However, the recombinants h^+r^+ and *hr* were detected in addition to the parental genotypes (see Figure 6.18). As with eukaryotes, the percentage of recombinant plaques divided by the total number of plaques reflects the relative distance between the genes:

$$\text{recombinational frequency} = (h^+r^+ + hr)/\text{total plaques} \times 100$$

Sample data for the *h* and *r* loci are shown in **Table 6.2**.

Similar recombinational studies have been conducted with numerous mutant genes in a variety of bacteriophages. Data are analyzed in much the same way as in eukaryotic mapping experiments. Two- and three-point mapping crosses are possible, and the percentage of recombinants in the total number of phage progeny is calculated. This value is proportional to the relative distance between two genes along the DNA molecule constituting the chromosome.

Investigations into phage recombination support a model similar to that of eukaryotic crossing over—a breakage and reunion process between the viral chromosomes. A fairly clear picture of the dynamics of viral recombination has emerged. Following the early phase of infection, the chromosomes of the phages begin replication. As this stage progresses, a pool of chromosomes accumulates in the bacterial cytoplasm. If double infection by phages of two genotypes has occurred, then the pool of chromosomes initially consists of the two parental types. Genetic exchange between these two types will occur before, during, and after replication, producing recombinant chromosomes.

In the case of the h^+r and hr^+ example discussed here, recombinant h^+r^+ and *hr* chromosomes are produced. Each of these chromosomes can undergo replication, with new replicates undergoing exchange with each other and with parental chromosomes. Furthermore, recombination is not

TABLE 6.1 Some Mutant Types of T-Even Phages

Name	Description
minute	Forms small plaques
turbid	Forms turbid plaques on *E. coli* B
star	Forms irregular plaques
UV-sensitive	Alters UV sensitivity
acriflavin-resistant	Forms plaques on acriflavin agar
osmotic shock	Withstands rapid dilution into distilled water
lysozyme	Does not produce lysozyme
amber	Grows on *E. coli* K12 but not B
temperature-sensitive	Grows at 25°C but not at 42°C

TABLE 6.2 Results of a Cross Involving the *h* and *r* Genes in Phage T2 ($hr^+ \times h^+r$)

Genotype	Plaques	Designation
hr^+	42	Parental progeny 76%
h^+r	34	
h^+r^+	12	Recombinants 24%
hr	12	

Source: Data derived from Hershey and Rotman (1949).

restricted to exchange between two chromosomes—three or more may be involved simultaneously. As phage development progresses, chromosomes are randomly removed from the pool and packed into the phage head, forming mature phage particles. Thus, a variety of parental and recombinant genotypes are represented in progeny phages.

As we will see in Section 6.8, powerful selection systems have made it possible to detect *intragenic* recombination in viruses, where exchanges occur at points within a single gene, as opposed to intergenic recombination, where exchanges occur at points located between genes. Such studies have led to what has been called the fine-structure analysis of the gene.

6.8 Intragenic Recombination Occurs in Phage T4

We conclude this chapter with an account of an ingenious example of genetic analysis. In the early 1950s, Seymour Benzer undertook a detailed examination of a single locus, *rII*, in phage T4. Benzer successfully designed experiments to recover the extremely rare genetic recombinants arising as a result of intragenic exchange. Such recombination is equivalent to eukaryotic crossing over, but in this case, within a gene rather than at a point between two genes. Benzer demonstrated that such recombination occurs between the DNA of individual bacteriophages during simultaneous infection of the host bacterium *E. coli*.

The end result of Benzer's work was the production of a detailed map of the *rII* locus. Because of the extremely detailed information provided by his analysis, and because these experiments occurred decades before DNA-sequencing techniques were developed, the insights concerning the internal structure of the gene were particularly noteworthy.

The *rII* Locus of Phage T4

The primary requirement in genetic analysis is the isolation of a large number of mutations in the gene being investigated. Mutants at the *rII* locus produce distinctive plaques when plated on *E. coli* strain B, allowing their easy identification. Figure 6.18 illustrates mutant *r* plaques compared to their wild-type r^+ counterparts in the related T2 phage. Benzer's approach was to isolate many independent *rII* mutants—he eventually obtained about 20,000—and to perform recombinational studies so as to produce a genetic map of this locus. Benzer assumed that most of these mutations, because they were randomly isolated, would represent different locations within the *rII* locus and would thus provide an ample basis for mapping studies.

The key to Benzer's analysis was that *rII* mutant phages, though capable of infecting and lysing *E. coli* B, could not successfully lyse a second related strain, *E. coli* K12(λ).* Wild-type phages, by contrast, could lyse both the B and the K12 strains. Benzer reasoned that these conditions provided the potential for a highly sensitive screening system. If phages from any two different mutant strains were allowed to simultaneously infect *E. coli* B, exchanges between the two mutant sites within the locus would produce rare wild-type recombinants (**Figure 6.19**). If the progeny phage population, which contained more than 99.9 percent *rII* phages and less than 0.1 percent wild-type phages, were then allowed to infect strain K12, the wild-type recombinants would successfully reproduce and produce wild-type plaques. *This is the critical step in recovering and quantifying rare recombinants.*

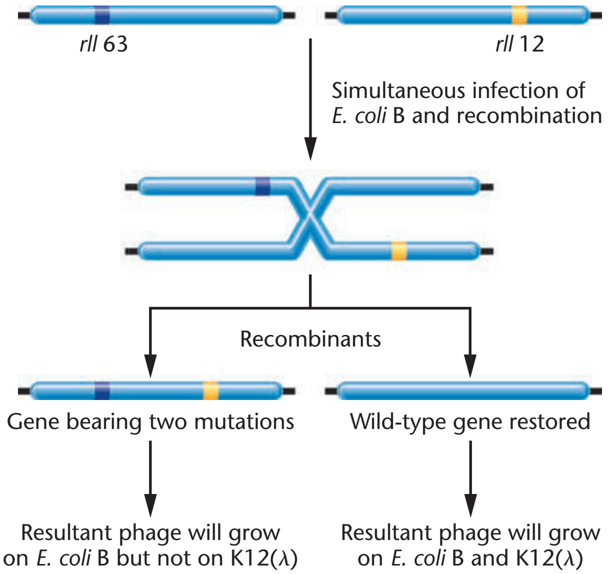

rII 63
rII 12
Simultaneous infection of E. coli B and recombination
Recombinants
Gene bearing two mutations
Wild-type gene restored
Resultant phage will grow on E. coli B but not on K12(λ)
Resultant phage will grow on E. coli B and K12(λ)

FIGURE 6.19 Illustration of intragenic recombination between two mutations in the *rII* locus of phage T4. The result is the production of a wild-type phage, which will grow on both *E. coli* B and K12, and of a phage that has incorporated both mutations into the *rII* locus. The latter will grow on *E. coli* B but not on *E. coli* K12.

By using serial dilution techniques, Benzer was able to determine the total number of mutant *rII* phages produced on *E. coli* B and the total number of recombinant wild-type phages that would lyse *E. coli* K12. These data provided the basis for calculating the frequency of recombination, a value proportional to the distance within the gene between the two mutations being studied. As we will see, this experimental design was extraordinarily sensitive. Remarkably, it was possible for Benzer to detect as few as one recombinant wild-type phage among 100 million mutant phages.

* The inclusion of "(λ)" in the designation of K12 indicates that this bacterial strain is lysogenized by phage λ. This, in fact, is the reason that *rII* mutants cannot lyse such bacteria. In future discussions, this strain will simply be abbreviated as *E. coli* K12.

When information from many such experiments is combined, a detailed map of the locus is possible.

Before we discuss this mapping, we need to describe an important discovery Benzer made during the early development of his screen—a discovery that led to the development of a technique used widely in genetics labs today, the **complementation assay** you learned about earlier (see Chapter 4).

Complementation by *rII* Mutations

Before Benzer was able to initiate these intragenic recombination studies, he had to resolve a problem encountered during the early stages of his experimentation. While doing a control study in which K12 bacteria were simultaneously infected with pairs of different *rII* mutant strains, Benzer sometimes found that certain pairs of the *rII* mutant strains lysed the K12 bacteria. This was initially quite puzzling, since only the wild-type *rII* was supposed to be capable of lysing K12 bacteria. How could two mutant strains of *rII*, each of which was thought to contain a defect in the same gene, show a wild-type function?

Benzer reasoned that, during simultaneous infection, each mutant strain provided something that the other lacked, thus restoring wild-type function. This phenomenon, which he called **complementation,** is illustrated in Figure 6.20(a). When many pairs of mutations were tested, each mutation fell into one of two possible **complementation groups,** A or B. Those that failed to complement one another were placed in the same complementation group, while those that did complement one another were each assigned to a different complementation group. Benzer coined the term **cistron,** which he defined as the smallest functional genetic unit, to describe a complementation group. In modern terminology, we know that a cistron represents a gene.

We now know that Benzer's A and B cistrons represent two separate genes in what we originally referred to as the *rII* locus (because of the initial assumption that it was a single gene). Complementation occurs when K12 bacteria are infected with two *rII* mutants, one with a mutation in the A gene and one with a mutation in the B gene. Therefore, there is a source of both wild-type gene products, since the A mutant provides wild-type B and the B mutant provides wild-type A. We can also explain why two strains that fail to complement, say two A-cistron mutants, are actually mutations in the same gene. In this case, if two A-cistron mutants are combined, there will be an immediate source of the wild-type B product, but no immediate source of the wild-type A product [Figure 6.20(b)].

Once Benzer was able to place all *rII* mutations in either the A or the B cistron, he was set to return to his intragenic recombination studies, testing mutations in the A cistron against each other and testing mutations in the B cistron against each other.

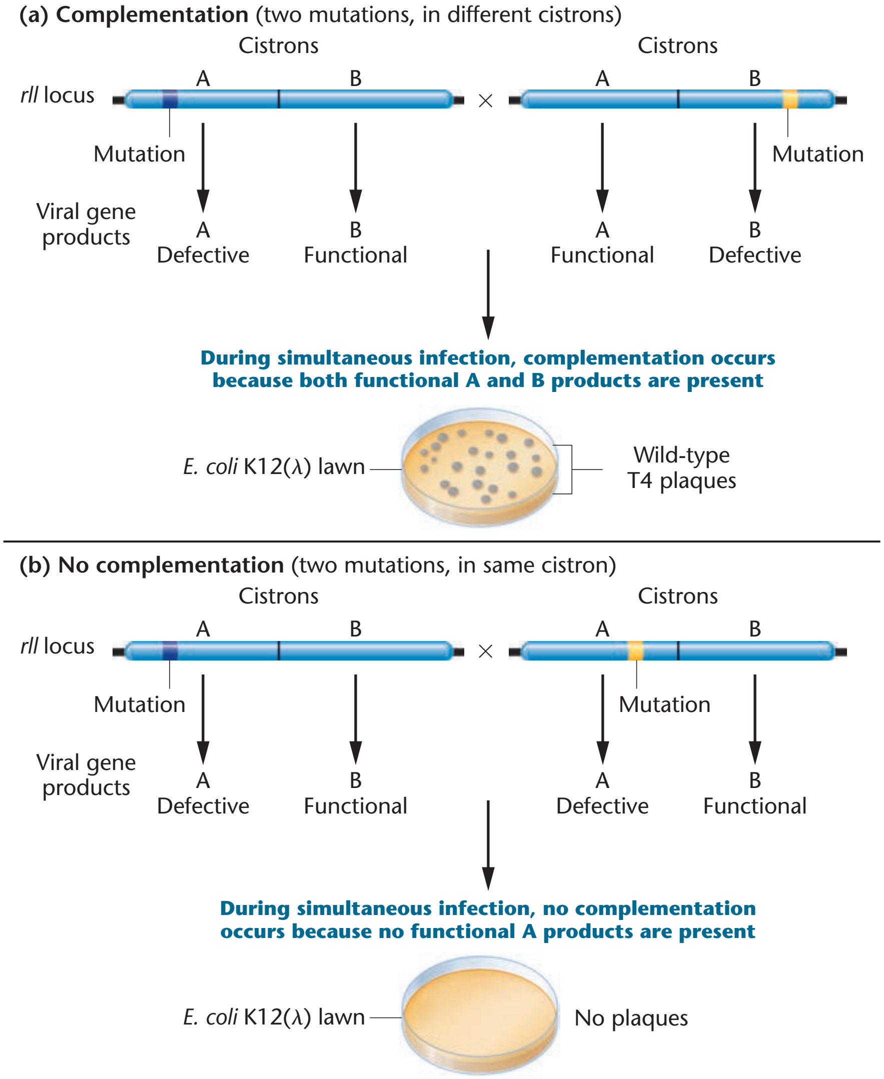

(a) Complementation (two mutations, in different cistrons)
Cistrons
Cistrons
rll locus
A
B
A
B
×
Mutation
Mutation
Viral gene products
A Defective
B Functional
A Functional
B Defective
During simultaneous infection, complementation occurs because both functional A and B products are present
E. coli K12(λ) lawn
Wild-type T4 plaques
(b) No complementation (two mutations, in same cistron)
Cistrons
Cistrons
rll locus
A
B
A
B
×
Mutation
Mutation
Viral gene products
A Defective
B Functional
A Defective
B Functional
During simultaneous infection, no complementation occurs because no functional A products are present
E. coli K12(λ) lawn
No plaques

FIGURE 6.20 Comparison of two pairs of *rII* mutations. (a) In one case, they complement one another. (b) In the other case, they do not complement one another. Complementation occurs when each mutation is in a separate cistron. Failure to complement occurs when the two mutations are in the same cistron.

NOW SOLVE THIS

6.3 In complementation studies of the *rII* locus of phage T4, three groups of three different mutations were tested. For each group, only two combinations were tested. On the basis of each set of data (shown here), predict the results of the third experiment for each group.

Group A	Group B	Group C
d × *e*–lysis	*g* × *b*–no lysis	*j* × *k*–lysis
d × *f*–no lysis	*g* × *i*–no lysis	*j* × *l*–lysis
e × *f*–?	*b* × *i*–?	*k* × *l*–?

■ **HINT:** *This problem involves an understanding of why complementation occurs during simultaneous infection of a bacterial cell by two bacteriophage strains, each with a different mutation within the* rII *locus. The key to its solution is to be aware that if each mutation alters a different genetic product, then each strain will provide the product that the other is missing, thus leading to complementation.*

For more practice, see Problems 18, 19, and 21.

Recombinational Analysis

Of the approximately 20,000 *rII* mutations, roughly half fell into each cistron. Benzer set about mapping the mutations within each one. For example, if two *rII* A mutants (i.e., two phage strains with different mutations in the A cistron) were first allowed to infect *E. coli* B in a liquid culture, and if a recombination event occurred between the mutational sites in the A cistron, then wild-type progeny viruses would be produced at low frequency. If samples of the progeny viruses from such an experiment were then plated on *E. coli* K12, only the wild-type recombinants would lyse the bacteria and produce plaques. The total number of nonrecombinant progeny viruses would be determined by plating samples on *E. coli* B.

This experimental protocol is illustrated in **Figure 6.21**. The percentage of recombinants can be determined by counting the plaques at the appropriate dilution in each case. As in eukaryotic mapping experiments, the frequency of recombination is an estimate of the distance between the two mutations within the cistron. For example, if the number of recombinants is equal to 4×10^3/mL, and the total number of progeny is 8×10^9/mL, then the frequency of recombination between the two mutants is

$$\begin{aligned} 2\left(\frac{4 \times 10^3}{8 \times 10^9}\right) &= 2(0.5 \times 10^{-6}) \\ &= 10^{-6} \\ &= 0.000001 \end{aligned}$$

Multiplying by 2 is necessary because each recombinant event yields two reciprocal products, only one of which—the wild type—is detected.

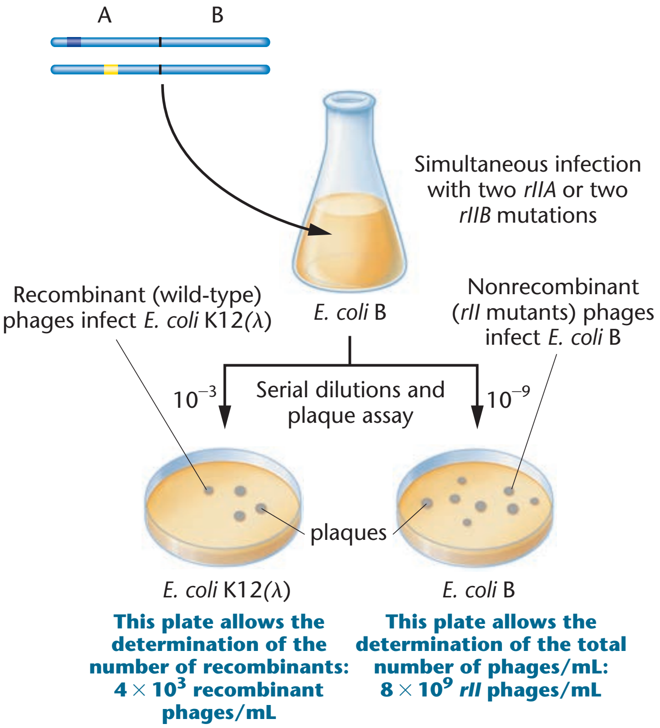

A
B
Simultaneous infection with two *rIIA* or two *rIIB* mutations
E. coli B
Recombinant (wild-type) phages infect *E. coli* K12(λ)
Nonrecombinant (*rII* mutants) phages infect *E. coli* B
10^{-3}
Serial dilutions and plaque assay
10^{-9}
plaques
E. coli K12(λ)
E. coli B
This plate allows the determination of the number of recombinants: 4×10^3 recombinant phages/mL
This plate allows the determination of the total number of phages/mL: 8×10^9 *rII* phages/mL

FIGURE 6.21 The experimental protocol for recombination studies between pairs of mutations in the same cistron. In this figure, all phage infecting *E. coli* B (in the flask) contain one of two mutations in the A cistron, as shown in the depiction of their chromosomes to the left of the flask.

Deletion Testing of the *rII* Locus

Although the system for assessing recombination frequencies described earlier allowed for mapping mutations within each cistron, testing 1000 mutants two at a time in all combinations would have required millions of experiments. Fortunately, Benzer was able to overcome this obstacle when he devised an analytical approach referred to as **deletion testing.** He discovered that some of the *rII* mutations were, in reality, deletions of small parts of both cistrons. That is, the genetic changes giving rise to the *rII* properties were not a characteristic of point mutations. Most importantly, when a deletion mutation was tested using simultaneous infection by two phage strains, one having the deletion mutation and the other having a point mutation located in the deleted part of the same cistron, the test never yielded wild-type recombinants. The reason is illustrated in **Figure 6.22**. Because the deleted area is lacking the area of DNA containing the point mutation, no recombination is possible. Thus, a method was available that could roughly, but quickly, localize any mutation, provided it was contained within a region covered by a deletion.

Deletion testing could thus provide data for the initial localization of each mutation. For example, seven overlapping deletions spanning various regions of the A cistron

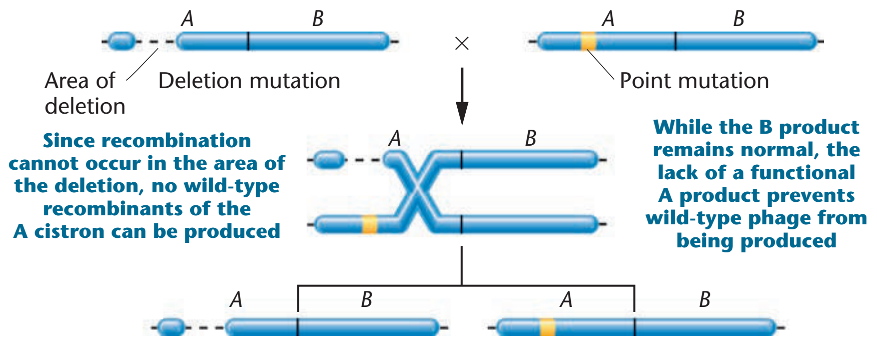

A
B
Area of deletion
Deletion mutation
×
A
B
Point mutation
Since recombination cannot occur in the area of the deletion, no wild-type recombinants of the A cistron can be produced
A
B
While the B product remains normal, the lack of a functional A product prevents wild-type phage from being produced
A
B
A
B

FIGURE 6.22 Demonstration that recombination between a phage chromosome with a deletion in the A cistron and another phage with a point mutation overlapped by that deletion cannot yield a chromosome with wild-type A and B cistrons.

were used for the initial screening of point mutations in that cistron, as shown in Figure 6.23. Depending on whether the viral chromosome bearing a point mutation does or does not undergo recombination with the chromosome bearing a deletion, each point mutation can be assigned to a specific area of the cistron. Further deletions within each of the seven areas can be used to localize, or map, each *rII* point mutation more precisely. Remember that, in each case, a point mutation is localized in the area of a deletion when it fails to give rise to any wild-type recombinants.

The *rII* Gene Map

After several years of work, Benzer produced a genetic map of the two cistrons composing the *rII* locus of phage T4 (Figure 6.24). From the 20,000 mutations analyzed, 307 distinct sites within this locus were mapped in relation to one another. Areas containing many mutations, designated as **hot spots,** were apparently more susceptible to mutation than were areas in which only one or a few mutations were found. In addition, Benzer discovered areas within the cistrons in which no mutations were localized. He estimated that as many as 200 recombinational units had not been localized by his studies.

EVOLVING CONCEPT OF A GENE

In the early 1950s, genes were regarded as indivisible units of heredity that could undergo mutation and between which *intergenic* recombination could occur. Seymour Benzer's pioneering genetic analysis of the *rII* locus in phage T4 in the mid-1950s demonstrated that the gene is not indivisible. Instead, he established that multiple sites exist within a gene, each capable of undergoing mutation, and between which *intragenic* recombination can occur. Benzer was able to map these sites within the *rII* locus. As we will see in an ensuing chapter, around the same time, it became clear that the gene is composed of DNA nucleotide pairs, each of which can be the site of mutation or recombination. ■

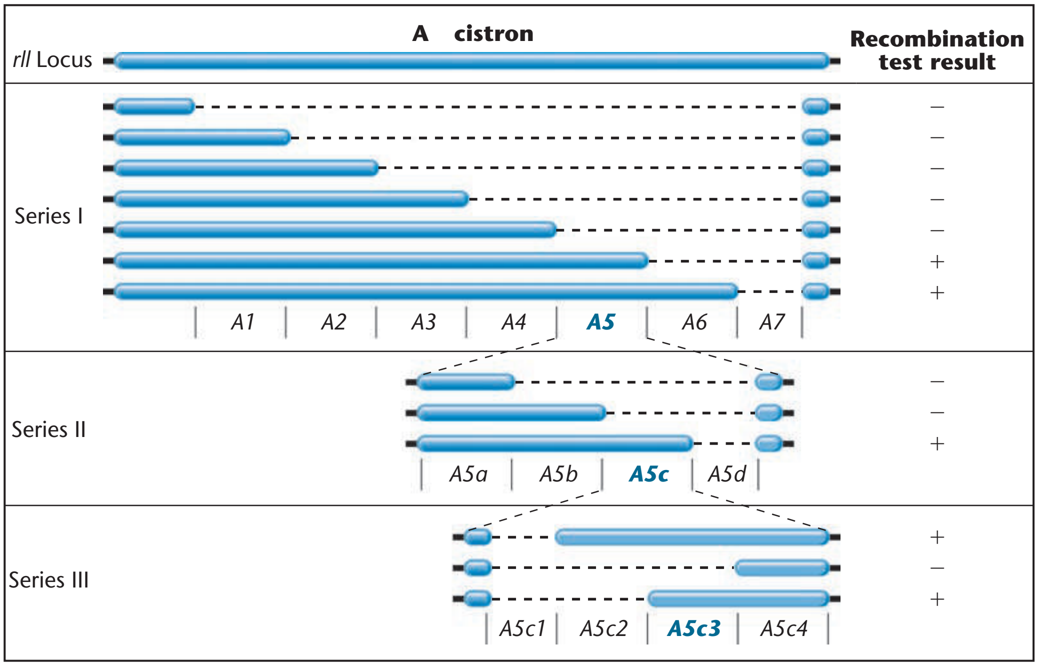


FIGURE 6.23 Three series of overlapping deletions in the A cistron of the *rII* locus used to localize the position of an unknown *rII* mutation. For example, if a mutant strain tested against each deletion (dashed areas) in series I for the production of recombinant wild-type progeny shows the results at the right (– or +), the mutation must be in segment *A5*. In series II, the mutation is further narrowed to segment *A5c*, and in series III to segment *A5c3*.

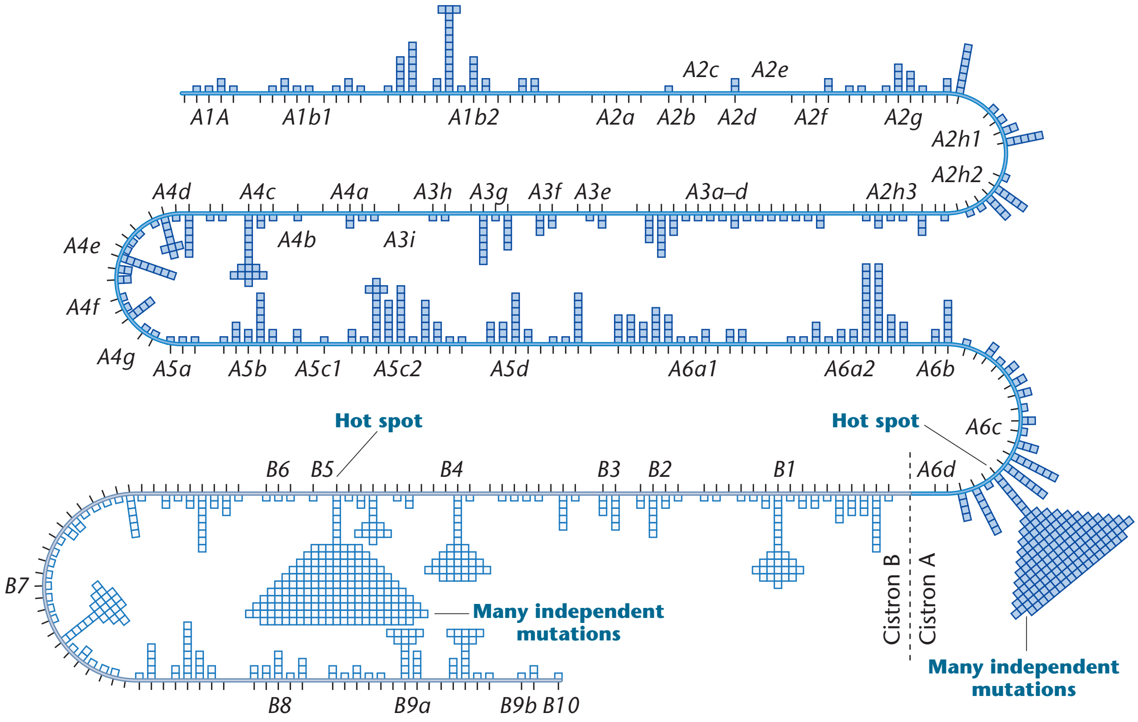


FIGURE 6.24 A partial map of mutations in the A and B cistrons of the *rII* locus of phage T4. Each square represents an independently isolated mutation. Note the two areas in which the largest number of mutations are present, referred to as "hot spots" (*A6cd* and *B5*).

GENETICS, ETHICS, AND SOCIETY

Multidrug-Resistant Bacteria: Fighting with Phage

The worldwide spread of *multidrug-resistant (MDR)* pathogenic bacteria has become an urgent threat to human and animal health. More than two million people in the United States become infected with antibiotic-resistant bacteria each year, and more than 23,000 of them will die from their infections. In 2015, approximately 480,000 cases of MDR tuberculosis occurred worldwide and another 100,000 cases were resistant to at least one antibiotic. In the United States, cases of drug-resistant enterobacteriaceae infections increased three-fold between 2001 and 2012. In 2016, a woman in Nevada died of a *Klebsiella pneumoniae* infection caused by a strain that was resistant to 26 different antibiotics, including colistin, which is considered the "last resort" antibiotic.

In an era when technology and science have made major gains in health diagnosis and treatment, why are we rapidly losing the battle against pathogenic bacteria? Can we find new antibiotics to replace those that are now ineffective? The answers to these questions are multifaceted and complicated by scientific, economic, and political issues.

One factor leading to the spread of MDR bacteria is the selective pressure brought about by repeated exposure to antibiotics. Worldwide, livestock consume as much as 80 percent of all antibiotics, used as feed supplements. The routine use of antibiotics in livestock feed and the overuse of human antibiotic prescriptions are thought to be the most significant contributors to the spread of MDR bacteria.

A second factor leading to the new "post-antibiotic era" is the reduction in antibiotic drug development by pharmaceutical companies. The fact that bacteria are developing resistance to so many drugs means that new and more difficult bacterial metabolic pathways need to be targeted. In addition, economic issues are significant. Drug companies spend hundreds of millions of dollars to develop and test a new drug. However, they receive less profit from antibiotics than from more expensive drugs such as chemotherapies or diabetes drugs.

Despite these challenges, several new antibiotic approaches are in development and early clinical trials. One of these approaches is the use of therapeutic bacteriophages (phages). Phages

(continued)

Genetics, Ethics, and Society, continued

have been used to treat bacterial infections since the early 1900s, especially in Europe, but were abandoned in the mid-twentieth century after the introduction of antibiotics. Researchers are returning to phage, using modern molecular tools to modify phage and phage-derived products for use as antibacterial drugs. No phage or phage products are yet approved for human therapies in the United States or Europe; however, several phage preparations, targeted at pathogens such as *Listeria*, are approved for topical use on fresh and prepared foods, and at least one phage therapy is in clinical trials.

Although scientific and regulatory challenges must still be overcome, we may be on the verge of the Age of the Phage.

Your Turn

Take time, individually or in groups, to consider the following questions. Investigate the references dealing with the technical and ethical challenges of combating drug-resistant bacteria.

1. How do phage therapies work, and what are the main advantages and disadvantages of using phage to treat bacterial infections?

These topics are discussed in Potera, C. (2013). Phage renaissance: new hope against antibiotic resistance (https://ehp.niehs.nih.gov/121-a48) *and* Cooper, C. J. et al. (2016). Adapting drug approval pathways for bacteriophage-based therapeutics. *Front. Microbiol*. 7:1–15.

2. Two significant reasons for the spread of MDR bacteria are the overuse of agricultural antibiotics and the reluctance of pharmaceutical companies to develop new antibiotics. Discuss the ethical concerns surrounded these two situations. For example, how do we balance our need for both abundant food and infection control? Also, how can we resolve the ethical disconnect between private-sector profits and the public good?

These topics are discussed in Littmann, J. and Viens, A. M. (2015). The ethical significance of antimicrobial resistance. *Public Health Ethics* 8:209–224.

CASE STUDY To treat or not to treat

A 4-month-old infant had been running a moderate fever for 36 hours, and a nervous mother made a call to her pediatrician. Examination and tests revealed no outward signs of infection or cause of the fever. The anxious mother wanted a prescription for antibiotics, but the pediatrician recommended watching the infant for two days before making a decision. He explained that decades of rampant use of antibiotics in medicine and agriculture has caused a global surge in antibiotic-resistant bacteria, drastically reducing the effectiveness of antibiotic therapy for infections. He pointed out that bacteria can exchange antibiotic resistance traits and that many pathogenic strains are now resistant to several antibiotics. The mother was not placated by these explanations and insisted that her baby receive antibiotics immediately. This situation raises several issues.

1. Was the pediatrician correct in stating that bacteria can exchange antibiotic resistance genes? If so, how is this possible?
2. If the infant was given antibiotics, how might this have contributed to the production of resistant bacteria?
3. If you were an anxious parent of the patient, would it change your mind if you learned that a woman died in 2016 from a bacterial infection that was resistant to all 26 antibiotics available in the United States?
4. How should the pediatrician balance his ethical responsibility to provide effective treatment to the present patient with his ethical responsibility to future patients who may need antibiotics for effective treatment?

See Garau, J. (2006). Impact of antibiotic restrictions: the ethical perspective. *Clin. Microbiol. Infect.* 12 (Supplement 5): 16–24. See also the Genetics, Ethics, and Society essay above.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Genetic recombination in bacteria takes place in three ways: conjugation, transformation, and transduction.
2. Conjugation may be initiated by a bacterium housing a plasmid called the F factor in its cytoplasm, making it a donor cell. Following conjugation, the recipient cell receives a copy of the F factor and is converted to the F^+ status.
3. When the F factor is integrated from the cytoplasm into the chromosome, the cell remains as a donor and is referred to as an Hfr cell. Upon mating, the donor chromosome moves unidirectionally into the recipient, initiating recombination and providing the basis for time mapping of the bacterial chromosome.
4. Plasmids, such as the F factor, are autonomously replicating DNA molecules found in the bacterial cytoplasm, sometimes containing unique genes conferring antibiotic resistance as well as the genes necessary for plasmid transfer during conjugation.
5. Transformation in bacteria, which does not require cell-to-cell contact, involves exogenous DNA that enters a recipient bacterium and recombines with the host's chromosome. Linkage mapping of closely aligned genes is possible during the analysis of transformation.
6. Bacteriophages, viruses that infect bacteria, demonstrate a well-defined life cycle where they reproduce within the host cell and can be studied using the plaque assay.

7. Bacteriophages can be lytic, meaning they infect the host cell, reproduce, and then lyse it, or in contrast, they can lysogenize the host cell, where they infect it and integrate their DNA into the host chromosome, but do not reproduce.
8. Transduction is virus-mediated bacterial DNA recombination. When a lysogenized bacterium subsequently reenters the lytic cycle, the new bacteriophages serve as vehicles for the transfer of host (bacterial) DNA.
9. Transduction is also used for bacterial linkage and mapping studies.
10. Various mutant phenotypes, including mutations in plaque morphology and host range, have been studied in bacteriophages. These have served as the basis for investigating genetic exchange and mapping in these viruses.
11. Genetic analysis of the *rII* locus in bacteriophage T4 allowed Seymour Benzer to study intragenic recombination. By isolating *rII* mutants and performing complementation analysis, recombinational studies, and deletion mapping, Benzer was able to locate and map more than 300 distinct sites within the two cistrons of the *rII* locus.

INSIGHTS AND SOLUTIONS

1. Time mapping is performed in a cross involving the genes *his, leu*, *mal*, and *xyl*. The recipient cells were auxotrophic for all four genes. After 25 minutes, mating was interrupted with the following results in recipient cells. Diagram the positions of these genes relative to the origin (*O*) of the F factor and to one another.

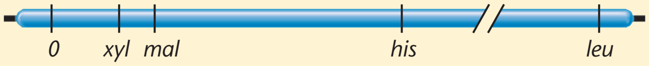
   (a) 90% were xyl^+

   (b) 80% were mal^+

   (c) 20% were his^+

   (d) none were leu^+

   Solution: The *xyl* gene was transferred most frequently, which shows it is closest to *O* (very close). The *mal* gene is next closest and reasonably near *xyl*, followed by the more distant *his* gene. The *leu* gene is far beyond these three, since no recombinants are recovered that include it. The diagram shows these relative locations along a piece of the circular chromosome.



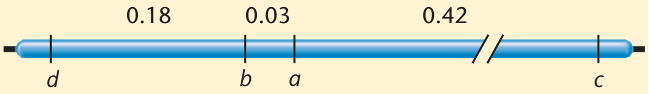
2. Three strains of bacteria, each bearing a separate mutation, a^-, b^-, or c^-, are the sources of donor DNA in a transformation experiment. Recipient cells are wild type for those genes but express the mutation d^-.

   (a) Based on the following data, and assuming that the location of the *d* gene precedes the *a, b*, and *c* genes, propose a linkage map for the four genes.

DNA Donor	Recipient	Transformants	Frequency of Transformants
a^-d^+	a^+d^-	a^+d^+	0.21
b^-d^+	b^+d^-	b^+d^+	0.18
c^-d^+	c^+d^-	c^+d^+	0.63

   (b) If the donor DNA were wild type and the recipient cells were either a^-b^-, a^-c^-, or b^-c^-, which of the crosses would be expected to produce the greatest number of wild-type transformants?

   Solution: (a) These data reflect the relative distances between the *a*, *b*, and *c* genes, individually, and the *d* gene. The *a* and *b* genes are about the same distance from the *d* gene and are thus tightly linked to one another. The *c* gene is more distant. Assuming that the *d* gene precedes the others, the map looks like this:



   (b) Because the *a* and *b* genes are closely linked, they most likely cotransform in a single event. Thus, recipient cells a^- b^- are most likely to convert to wild type.

3. For his fine-structure analysis of the *rII* locus in phage T4, Benzer was able to perform complementation testing of any pair of mutations once it was clear that the locus contained two cistrons. Complementation was assayed by simultaneously infecting *E. coli* K12 with two phage strains, each with an independent mutation, neither of which could alone lyse K12. From the data that follow, determine which mutations are in which cistron, assuming that mutation *1* (*M-1*) is in the A cistron and mutation *2* (*M-2*) is in the B cistron. Are there any cases where the mutation cannot be properly assigned?

Test Pair	Results*
1, 2	+
1, 3	–
1, 4	–
1, 5	+
2, 3	–
2, 4	+
2, 5	–

* + or – indicates complementation or the failure of complementation, respectively.

   Solution: *M-1* and *M-5* complement one another and, therefore, are not in the same cistron. Thus, *M-5* must be in the B cistron. *M-2* and *M-4* complement one another. By the same reasoning, *M-4* is not with *M-2* and, therefore, is in the A cistron. *M-3* fails to complement either *M-1* or *M-2*, and so it would seem to be in both cistrons. One explanation is that the physical cause of *M-3*

(continued)

Insights and Solutions—continued

somehow overlaps both the A and the B cistrons. It might be a double mutation with one sequence change in each cistron. It might also be a deletion that overlaps both cistrons and thus could not complement either *M-1* or *M-2*.

4. Another mutation, *M-6*, was tested with the results shown here:

Test Pair	Results
1, 6	+
2, 6	–
3, 6	–
4, 6	+
5, 6	–

Draw all possible conclusions about *M-6*.

Solution: These results are consistent with assigning *M-6* to the B cistron.

5. Recombination testing was then performed for *M-2*, *M-5*, and *M-6* so as to map the B cistron. Recombination analysis using both *E. coli* B and K12 showed that recombination occurred between *M-2* and *M-5* and between *M-5* and *M-6*, but not between *M-2* and *M-6*. Why not?

Solution: Either *M-2* and *M-6* represent identical mutations, or one of them may be a deletion that overlaps the other but does not overlap *M-5*. Furthermore, the data cannot rule out the possibility that both are deletions.

6. In recombination studies of the *rII* locus in phage T4, what is the significance of the value determined by calculating phage growth in the K12 versus the B strains of *E. coli* following simultaneous infection in *E. coli* B? Which value is always greater?

Solution: When plaque analysis is performed on *E. coli* B, in which the wild-type and mutant phages are both lytic, the total number of phages per milliliter can be determined. Because almost all cells are *rII* mutants of one type or another, this value is much larger than the value obtained with K12. To avoid total lysis of the plate, extensive dilution is necessary. In K12, *rII* mutations will not grow, but wild-type phages will. Because wild-type phages are the rare recombinants, there are relatively few of them and extensive dilution is not required.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we have focused on genetic systems present in bacteria and on the viruses that use bacteria as hosts (bacteriophages). In particular, we discussed mechanisms by which bacteria and their phages undergo genetic recombination, which allows geneticists to map bacterial and bacteriophage chromosomes. In the process, we found many opportunities to consider how this information was acquired. From the explanations given in the chapter, what answers would you propose to the following questions?
(a) How do we know that genes exist in bacteria and bacteriophages?
(b) How do we know that bacteria undergo genetic recombination, allowing the transfer of genes from one organism to another?
(c) How do we know whether or not genetic recombination between bacteria involves cell-to-cell contact?
(d) How do we know that bacteriophages recombine genetic material through transduction and that cell-to-cell contact is not essential for transduction to occur?
(e) How do we know that intergenic exchange occurs in bacteriophages?
(f) How do we know that in bacteriophage T4 the *rII* locus is subdivided into two regions, or cistrons?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p. 123. Many of these center around the findings that genetic recombination occurs in bacteria and in bacteriophages. Write a short summary that contrasts how recombination occurs in bacteria and bacteriophages.

3. With respect to F^+ and F^- bacterial matings, answer the following questions:
(a) How was it established that physical contact between cells was necessary?
(b) How was it established that chromosome transfer was unidirectional?
(c) What is the genetic basis for a bacterium's being F^+?

4. List all major differences between (a) the F^+ × F^- and the Hfr × F^- bacterial crosses; and (b) the F^+, F^-, Hfr, and F' bacteria.

5. Describe the basis for chromosome mapping in the Hfr × F^- crosses.

6. In general, when recombination experiments are conducted with bacteria, participating bacteria are mixed in complete medium, then transferred to a minimal growth medium. Why isn't the protocol reversed: minimal medium first, complete medium second?

7. Why are the recombinants produced from an Hfr × F^- cross rarely, if ever, F^+?

8. Describe the origin of F' bacteria and merozygotes.

9. In a transformation experiment, donor DNA was obtained from a prototroph bacterial strain ($a^+b^+c^+$), and the recipient was a triple auxotroph ($a^-b^-c^-$). What general conclusions can you draw about the linkage relationships among the three genes from the following transformant classes that were recovered?

$a^+\,b^-\,c^-$	180
$a^-\,b^+\,c^-$	150
$a^+\,b^+\,c^-$	210
$a^-\,b^-\,c^+$	179
$a^+\,b^-\,c^+$	2
$a^-\,b^+\,c^+$	1
$a^+\,b^+\,c^+$	3

10. Describe the role of heteroduplex formation during transformation.

11. Explain the observations that led Zinder and Lederberg to conclude that the prototrophs recovered in their transduction experiments were not the result of F^+ mediated conjugation.

12. Define plaque, lysogeny, and prophage.

13. Two theoretical genetic strains of a virus ($a^-b^-c^-$ and $a^+b^+c^+$) were used to simultaneously infect a culture of host bacteria. Of 10,000 plaques scored, the following genotypes were observed. Determine the genetic map of these three genes on the viral chromosome. Decide whether interference was positive or negative.

$a^+ b^+ c^+$	4100	$a^- b^+ c^-$	160
$a^- b^- c^-$	3990	$a^+ b^- c^+$	140
$a^+ b^- c^-$	740	$a^- b^- c^+$	90
$a^- b^+ c^+$	670	$a^+ b^+ c^-$	110

14. The bacteriophage genome consists of many genes encoding proteins that make up the head, collar, tail, and tail fibers. When these genes are transcribed following phage infection, how are these proteins synthesized, since the phage genome lacks genes essential to ribosome structure?

15. If a single bacteriophage infects one *E. coli* cell present on a lawn of bacteria and, upon lysis, yields 200 viable viruses, how many phages will exist in a single plaque if three more lytic cycles occur?

16. A phage-infected bacterial culture was subjected to a series of dilutions, and a plaque assay was performed in each case, with the results shown in the following table. What conclusion can be drawn in the case of each dilution, assuming that 0.1 mL was used in each plaque assay?

	Dilution Factor	Assay Results
(a)	10^4	All bacteria lysed
(b)	10^5	14 plaques
(c)	10^6	0 plaques

17. In recombination studies of the *rII* locus in phage T4, what is the significance of the value determined by calculating phage growth in the K12 versus the B strains of *E. coli* following simultaneous infection in *E. coli* B? Which value is always greater?

18. In an analysis of *rII* mutants, complementation testing yielded the following results:

Mutants	Results (+/− lysis)
1, 2	+
1, 3	+
1, 4	−
1, 5	−

Predict the results of testing *2* and *3*, *2* and *4*, and *3* and *4* together.

19. If further testing of the mutations in Problem 18 yielded the following results, what would you conclude about mutant *5*?

Mutants	Results
2, 5	−
3, 5	−
4, 5	−

20. Using mutants *2* and *3* from Problem 19, following mixed infection on *E. coli* B, progeny viruses were plated in a series of dilutions on both *E. coli* B and K12 with the following results.
(a) What is the recombination frequency between the two mutants?

Strain Plated	Dilution	Plaques
E. coli B	10^{-5}	2
E. coli K12	10^{-1}	5

(b) Another mutation, *6*, was tested in relation to mutations *1* through *5* from Problems 18–20. In initial testing, mutant *6* complemented mutants *2* and *3*. In recombination testing with *1*, *4*, and *5*, mutant *6* yielded recombinants with *1* and *5*, but not with *4*. What can you conclude about mutation 6?

Extra-Spicy Problems

21. During the analysis of seven *rII* mutations in phage T4, mutants *1, 2,* and *6* were in cistron A, while mutants *3, 4,* and *5* were in cistron B. Of these, mutant *4* was a deletion overlapping mutant *5*. The remainder were point mutations. Nothing was known about mutant *7*. Predict the results of complementation (+ or −) between *1* and *2; 1* and *3; 2* and *4*; and *4* and *5*.

22. In studies of recombination between mutants *1* and *2* from Problem 21, the results shown in the following table were obtained.

Strain	Dilution	Plaques	Phenotypes
E. coli B	10^{-7}	4	*r*
E. coli K12	10^{-2}	8	+

(a) Calculate the recombination frequency.

(b) When mutant *6* was tested for recombination with mutant *1*, the data were the same as those shown above for strain B, but not for K12. The researcher lost the K12 data, but remembered that recombination was ten times more frequent than when mutants *1* and *2* were tested. What were the lost values (dilution and plaque numbers)?

(c) Mutant *7* (Problem 21) failed to complement any of the other mutants (*1*–*6*). Define the nature of mutant *7*.

23. In *Bacillus subtilis*, linkage analysis of two mutant genes affecting the synthesis of two amino acids, tryptophan (trp_2^-) and tyrosine (tyr_1^-), was performed using transformation. Examine the following data and draw all possible conclusions regarding linkage. What is the purpose of Part B of the experiment? [Reference: E. Nester, M. Schafer, and J. Lederberg (1963).]

Donor DNA	Recipient Cell	Transformants	No.
A. trp_2^+ tyr_1^+	trp_2^- tyr_1^-	trp^+ tyr^-	196
		trp^- tyr^+	328
		trp^+ tyr^+	367
trp_2^+ tyr_1^-	trp_2^- tyr_1^-	trp^+ tyr^-	190
B. and		trp^- tyr^+	256
trp_2^- tyr_1^+		trp^+ tyr^+	2

24. An Hfr strain is used to map three genes in an interrupted mating experiment. The cross is $Hfr/a^+b^+c^+rif \times F^-/a^-b^-c^-rif^r$. (No map order is implied in the listing of the alleles; rif^r is resistance to the antibiotic rifampicin.) The a^+ gene is required for the biosynthesis of nutrient A, the b^+ gene for nutrient B, and c^+ for nutrient C. The minus alleles are auxotrophs for these nutrients. The cross is initiated at time = 0, and at various times, the mating mixture is plated on three types of medium. Each plate contains minimal medium (MM) plus rifampicin plus specific supplements that are indicated in the following table. (The results for each time interval are shown as the number of colonies growing on each plate.)

	Time of Interruption			
	5 min	10 min	15 min	20 min
Nutrients A and B	0	0	4	21
Nutrients B and C	0	5	23	40
Nutrients A and C	4	25	60	82

(a) What is the purpose of rifampicin in the experiment?

(b) Based on these data, determine the approximate location on the chromosome of the *a*, *b*, and *c* genes relative to one another and to the F factor.

(c) Can the location of the *rif* gene be determined in this experiment? If not, design an experiment to determine the location of *rif* relative to the F factor and to gene *b*.

25. A plaque assay is performed beginning with 1 mL of a solution containing bacteriophages. This solution is serially diluted three times by combining 0.1 mL of each sequential dilution with 9.9 mL of liquid medium. Then 0.1 mL of the final dilution is plated in the plaque assay and yields 17 plaques. What is the initial density of bacteriophages in the original 1 mL?

26. In a cotransformation experiment, using various combinations of genes two at a time, the following data were produced. Determine which genes are "linked" to which others.

Successful Cotransformation	Unsuccessful Cotransformation
a and *d*; *b* and *c*;	*a* and *b*; *a* and *c*; *a* and *f*;
b and *f*	*d* and *b*; *d* and *c*; *d* and *f*;
	a and *e*; *b* and *e*; *c* and *e*;
	d and *e*; *f* and *e*

27. For the experiment in Problem 26, another gene, *g*, was studied. It demonstrated positive cotransformation when tested with gene *f*. Predict the results of testing gene *g* with genes *a*, *b*, *c*, *d*, and *e*.

28. Bacterial conjugation, mediated mainly by conjugative plasmids such as F, represents a potential health threat through the sharing of genes for pathogenicity or antibiotic resistance. Given that more than 400 different species of bacteria coinhabit a healthy human gut and more than 200 coinhabit human skin, Francisco Dionisio [(2002) *Genetics* 162:1525–1532] investigated the ability of plasmids to undergo between-species conjugal transfer. The following data are presented for various species of the enterobacterial genus *Escherichia*. The data are presented as "log base 10" values; for example, −2.0 would be equivalent to 10^{-2} as a rate of transfer. Assume that all differences between values presented are statistically significant.

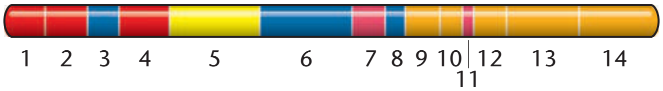
(a) What general conclusion(s) can be drawn from these data?

(b) In what species is within-species transfer most likely? In what species pair is between-species transfer most likely?

(c) What is the significance of these findings in terms of human health?

	Donor			
Recipient	*E. chrysanthemi*	*E. blattae*	*E. fergusonii*	*E. coli*
E. chrysanthemi	−2.4	−4.7	−5.8	−3.7
E. blattae	−2.0	−3.4	−5.2	−3.4
E. fergusonii	−3.4	−5.0	−5.8	−4.2
E. coli	−1.7	−3.7	−5.3	−3.5

29. A study was conducted in an attempt to determine which functional regions of a particular conjugative transfer gene (*tra1*) are involved in the transfer of plasmid R27 in *Salmonella enterica*. The R27 plasmid is of significant clinical interest because it is capable of encoding multiple-antibiotic resistance to typhoid fever. To identify functional regions responsible for conjugal transfer, an analysis by Lawley et al. [(2002). *J. Bacteriol.* 184:2173–2180] was conducted in which particular regions of the *tra1* gene were mutated and tested for their impact on conjugation. Shown here is a map of the regions tested and believed to be involved in conjugative transfer of the plasmid. Similar coloring indicates related function. Numbers correspond to each functional region subjected to mutation analysis.



Accompanying the map is a table showing the effects of these mutations on R27 conjugation.

Effects of Mutations in Functional Regions of Transfer Region 1 (*tra1*) on R27 Conjugation

R27 Mutation in Region	Conjugative Transfer	Relative Conjugation Frequency (%)
1	+	100
2	+	100
3	−	0
4	+	100
5	−	0
6	−	0
7	+	12
8	−	0
9	−	0
10	−	0
11	+	13
12	−	0
13	−	0
14	−	0

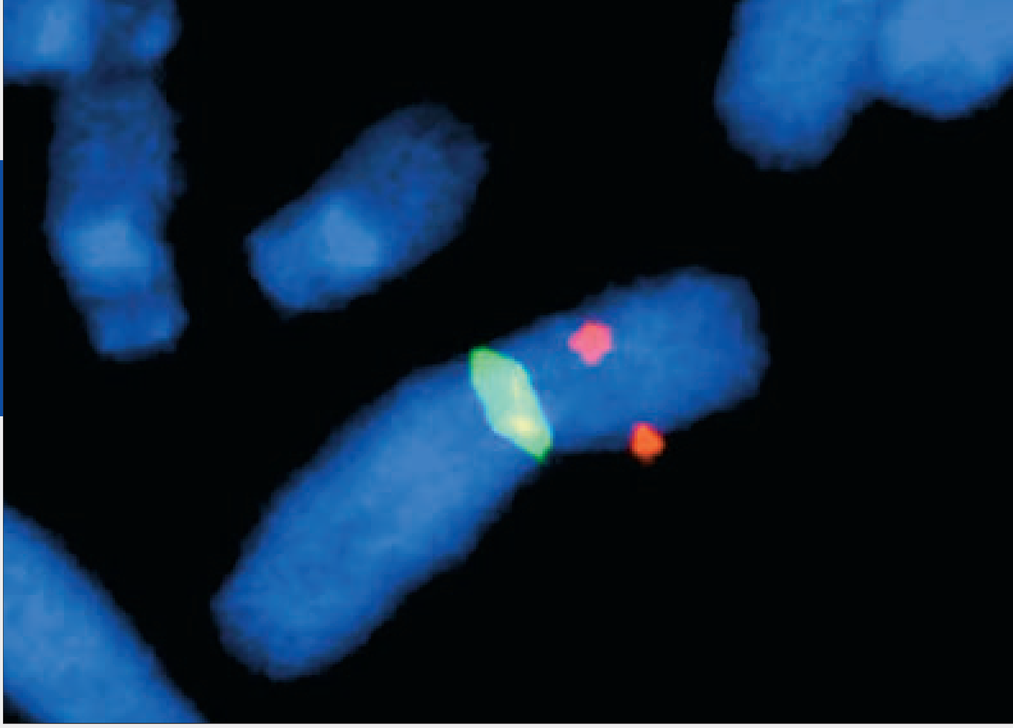
(a) Given the data, do all functional regions appear to influence conjugative transfer?

(b) Which regions appear to have the most impact on conjugation?

(c) Which regions appear to have a limited impact on conjugation?

(d) What general conclusions might one draw from these data?

7



Sex Determination and Sex Chromosomes

A human X chromosome highlighted using fluorescence *in situ* hybridization (FISH), a method in which specific probes bind to specific sequences of DNA. The green fluorescence probe binds to DNA at the centromere of X chromosomes. The red fluorescence probe binds to the DNA sequence of the X-linked Duchenne muscular dystrophy (DMD) gene.

CHAPTER CONCEPTS

- A variety of mechanisms have evolved that result in sexual differentiation, leading to sexual dimorphism and greatly enhancing the production of genetic variation within species.
- Often, specific genes, usually on a single chromosome, cause maleness or femaleness during development.
- In humans, the presence of extra X or Y chromosomes beyond the diploid number may be tolerated but often leads to syndromes demonstrating distinctive phenotypes.
- While segregation of sex-determining chromosomes should theoretically lead to a one-to-one sex ratio of males to females, in humans the actual ratio favors males at conception.
- In mammals, females inherit two X chromosomes compared to one in males, but the extra genetic information in females is compensated for by random inactivation of one of the X chromosomes early in development.
- In some reptilian species, temperature during incubation of eggs determines the sex of offspring.

In the biological world, a wide range of reproductive modes and life cycles are observed. Some organisms are entirely asexual, displaying no evidence of sexual reproduction. Other organisms alternate between short periods of sexual reproduction and prolonged periods of asexual reproduction. In most diploid eukaryotes, however, sexual reproduction is the only natural mechanism for producing new members of the species. The perpetuation of all sexually reproducing organisms depends ultimately on an efficient union of gametes during fertilization. In turn, successful fertilization depends on some form of **sexual differentiation** in the reproductive organisms. Even though it is not overtly evident, this differentiation occurs in organisms as low on the evolutionary scale as bacteria and single-celled eukaryotic algae. In more complex forms of life, the differentiation of the sexes is more evident as phenotypic dimorphism of males and females. The ancient symbol for iron and for Mars, depicting a shield and spear (♂), and the ancient symbol for copper and for Venus, depicting a mirror (♀), have also come to symbolize maleness and femaleness, respectively.

Dissimilar, or **heteromorphic**, **chromosomes**, such as the XY pair in mammals, characterize one sex or the other in a wide range of species, resulting in their label as **sex chromosomes.** Nevertheless, it is genes, rather than chromosomes, that ultimately serve as the underlying basis of **sex determination**. As we will see, some of these genes are present on sex chromosomes, but others are autosomal. Extensive investigation has revealed a wide variation in sex-chromosome systems—even in closely related organisms—suggesting that mechanisms controlling sex determination have undergone rapid evolution many times in the history of life.

In this chapter, we delve more deeply into what is known about the genetic basis for the determination of sexual differences, with a particular emphasis on two organisms: our own species, representative of mammals; and *Drosophila*, on which pioneering sex-determining studies were performed.

7.1 X and Y Chromosomes Were First Linked to Sex Determination Early in the Twentieth Century

How sex is determined has long intrigued geneticists. In 1891, Hermann Henking identified a nuclear structure in the sperm of certain insects, which he labeled the X-body. Several years later, Clarence McClung showed that some of the sperm in grasshoppers contain an unusual genetic structure, called a *heterochromosome,* but the remainder of the sperm lack this structure. He mistakenly associated the presence of the heterochromosome with the production of male progeny. In 1906, Edmund B. Wilson clarified Henking and McClung's findings when he demonstrated that female somatic cells in the butterfly *Protenor* contain 14 chromosomes, including 12 autosomes (A) and two X chromosomes. During oogenesis, an even reduction occurs, producing gametes with seven chromosomes, including one X chromosome (6A + X). Male somatic cells, on the other hand, contain only 13 chromosomes, including one X chromosome. During spermatogenesis, gametes are produced containing either six chromosomes, without an X (6A), or seven chromosomes, one of which is an X (6A + X). Fertilization by X-bearing sperm results in female offspring, and fertilization by X-deficient sperm results in male offspring [Figure 7.1(a)].

The presence or absence of the X chromosome in male gametes provides an efficient mechanism for sex determination in this species and also produces a 1:1 sex ratio in the resulting offspring.

Wilson also experimented with the milkweed bug *Lygaeus turcicus*, in which both sexes have 14 chromosomes. Twelve of these are autosomes. In addition, the females have two X chromosomes, while the males have only a single X and a smaller heterochromosome labeled the **Y chromosome.** Females in this species produce only gametes of the (6A + X) constitution, but males produce two types of gametes in equal proportions, (6A + X) and (6A + Y). Therefore, following random fertilization, equal numbers of male and female progeny will be produced with distinct chromosome complements [Figure 7.1(b)].

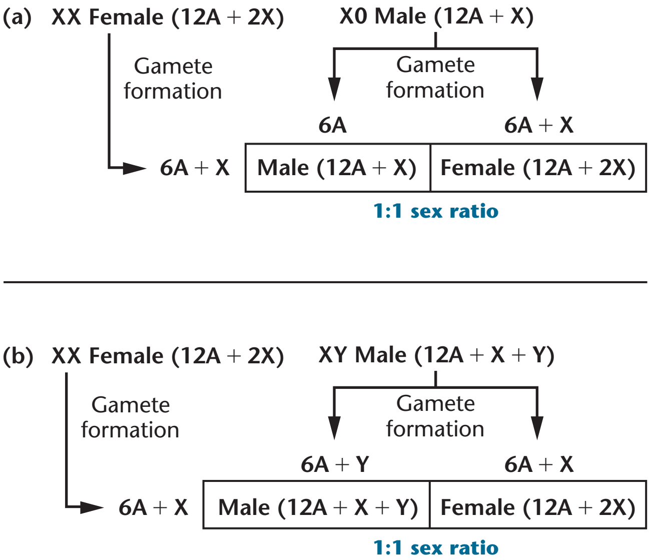


FIGURE 7.1 (a) Sex determination where the heterogametic sex (the male in this example) is X0 and produces gametes with or without the X chromosome; (b) sex determination, where the heterogametic sex (again, the male in this example) is XY and produces gametes with either an X or a Y chromosome. In both cases, the chromosome composition of the offspring determines its sex.

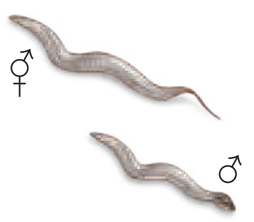
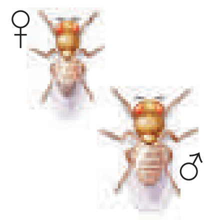
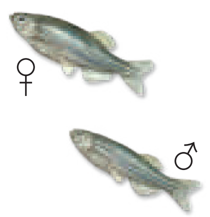
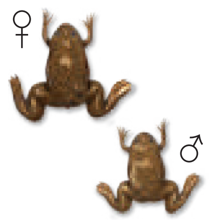
In *Protenor* and *Lygaeus*, males produce gametes with different chromosome compositions. As a result, they are described as the **heterogametic sex,** and in effect, their gametes ultimately determine the sex of the progeny in those species. In such cases, the female, which has like sex chromosomes, is the **homogametic sex,** producing uniform gametes with regard to chromosome numbers and types.

The male is not always the heterogametic sex. In some organisms, the female produces unlike gametes, exhibiting either the *Protenor* XX/XO or *Lygaeus* XX/XY mode of sex determination. Examples include certain moths and butterflies, some fish, reptiles, amphibians, at least one species of plants (*Fragaria orientalis*), and most birds. To immediately distinguish situations in which the female is the heterogametic sex, some geneticists use the notation **ZZ/ZW,** where ZZ is the homogametic male and ZW is the heterogametic female, instead of the XX/XY notation. For example, chickens are so denoted. The sex chromosome composition for popular model organisms in genetics is shown in Table 7.1.

7.2 The Y Chromosome Determines Maleness in Humans

The first attempt to understand sex determination in our own species occurred almost 100 years ago and involved the visual examination of chromosomes in dividing cells.

TABLE 7.1 Sex Chromosome Compositions of Common Model Organisms

	Caenorhabditis elegans	*Drosophila melanogaster*	*Mus musculus*	*Danio rerio*	*Xenopus laevis*
Model Organism					
Sex Chromosomes	XX XO	XX XY	XX XY	None	ZW ZZ

Efforts were made to accurately determine the diploid chromosome number of humans, but because of the relatively large number of chromosomes, this proved to be quite difficult. Then, in 1956, Joe Hin Tjio and Albert Levan discovered an effective way to prepare chromosomes for accurate viewing. This technique led to a strikingly clear demonstration of metaphase stages showing that 46 was indeed the human diploid number. Later that same year, C. E. Ford and John L. Hamerton, also working with testicular tissue, confirmed this finding. The familiar karyotypes of a human male (Figure 2.4) illustrate the difference in size between the human X and Y chromosomes.

Of the normal 23 pairs of human chromosomes, one pair was shown to vary in configuration in males and females. These two chromosomes were designated the X and Y sex chromosomes. The human female has two X chromosomes, and the human male has one X and one Y chromosome.

We might believe that this observation is sufficient to conclude that the Y chromosome determines maleness. However, several other interpretations are possible. The Y could play no role in sex determination; the presence of two X chromosomes could cause femaleness; or maleness could result from the lack of a second X chromosome. The evidence that clarified which explanation was correct came from study of the effects of human sex-chromosome variations, described in the following section. As such investigations revealed, the Y chromosome does indeed determine maleness in humans.

Klinefelter and Turner Syndromes

Around 1940, scientists identified two human abnormalities characterized by aberrant sexual development, **Klinefelter syndrome (47,XXY)** and **Turner syndrome (45,X).*** Individuals with Klinefelter syndrome are generally tall and have long arms and legs and large hands and feet. They usually have genitalia and internal ducts that are male, but their testes are rudimentary and fail to produce sperm. At the same time, feminine sexual development is not entirely suppressed. Slight enlargement of the breasts (gynecomastia) is common, and the hips are often rounded. This ambiguous sexual development, referred to as intersexuality, can lead to abnormal social development. Intelligence is often below the normal range as well.

In Turner syndrome, the affected individual has female external genitalia and internal ducts, but the ovaries are rudimentary. Other characteristic abnormalities include short stature (usually under 5 feet), skin flaps on the back of the neck, and underdeveloped breasts. A broad, shieldlike chest is sometimes noted. Intelligence is usually normal.

In 1959, the karyotypes of individuals with these syndromes were determined to be abnormal with respect to the sex chromosomes. Individuals with Klinefelter syndrome have more than one X chromosome. Most often they have an XXY complement in addition to 44 autosomes [**Figure 7.2(a)**], which is why people with this karyotype are designated 47,XXY. Individuals with Turner syndrome most often have only 45 chromosomes, including just a single X chromosome; thus, they are designated 45,X [**Figure 7.2(b)**]. Note the convention used in designating these chromosome compositions. The number states the total number of chromosomes present, and the information after the comma indicates the deviation from the normal diploid content. Both conditions result from **nondisjunction,** the failure of the sex chromosomes to segregate properly during meiosis (nondisjunction is described in Chapter 8 and illustrated in Figure 8.1).

* Although the possessive form of the names of eponymous syndromes is sometimes used (e.g., Klinefelter's syndrome), the current preference is to use the nonpossessive form.

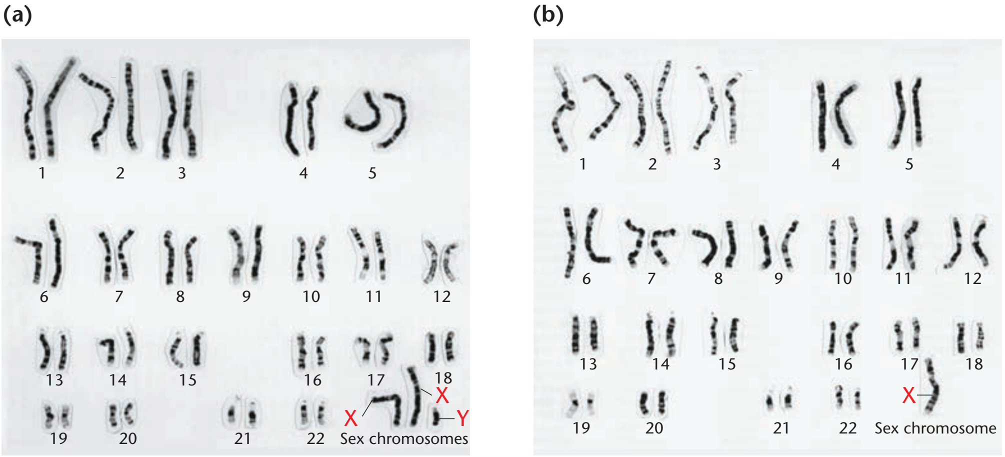


FIGURE 7.2 The karyotypes of individuals with (a) Klinefelter syndrome (47,XXY) and (b) Turner syndrome (45,X).

These Klinefelter and Turner karyotypes and their corresponding sexual phenotypes led scientists to conclude that the Y chromosome determines maleness in humans. In its absence, the person's sex is female, even if only a single X chromosome is present. The presence of the Y chromosome in the individual with Klinefelter syndrome is sufficient to determine maleness, even though male development is not complete. Similarly, in the absence of a Y chromosome, as in the case of individuals with Turner syndrome, no masculinization occurs. Note that we cannot conclude anything regarding sex determination under circumstances where a Y chromosome is present without an X because Y-containing human embryos lacking an X chromosome (designated 45,Y) do not survive.

Klinefelter syndrome occurs in about 1 of every 660 male births and is the most common sex chromosome disorder in males. The karyotypes **48,XXXY, 48,XXYY, 49,XXXXY**, and **49,XXXYY** are similar phenotypically to 47,XXY, but manifestations are often more severe in individuals with a greater number of X chromosomes. Recent studies have also shown that the variability in phenotypes for men with a 47,XXY genotype is correlated with copy number variations (CNVs), particularly duplications, on the X chromosomes.

Turner syndrome can also result from karyotypes other than 45,X, including individuals called **mosaics,** whose somatic cells display two different genetic cell lines, each exhibiting a different karyotype. Such cell lines result from a mitotic error during early development, the most common chromosome combinations being **45,X/46,XY** and **45,X/46,XX.** Thus, an embryo that began life with a normal karyotype can give rise to an individual whose cells show a mixture of karyotypes and who exhibits varying aspects of this syndrome.

Turner syndrome is observed in about 1 in 2000 female births, a frequency much lower than that for Klinefelter syndrome. One explanation for this difference is the observation that a substantial majority of 45,X fetuses die *in utero* and are aborted spontaneously. Thus, a similar frequency of the two syndromes may occur at conception.

47,XXX Syndrome

The abnormal presence of three X chromosomes along with a normal set of autosomes (**47,XXX**) results in female differentiation. The highly variable syndrome that accompanies this genotype, often called **triplo-X,** occurs in about 1 of 1000 female births. Frequently, 47,XXX women are perfectly normal and may remain unaware of their abnormality in chromosome number unless a karyotype is done. In other cases, underdeveloped secondary sex characteristics, sterility, delayed development of language and motor skills, and mental retardation may occur. In rare instances, **48,XXXX** (tetra-X) and **49,XXXXX** (penta-X) karyotypes have been reported. The syndromes associated with these karyotypes are similar to but more pronounced than the 47,XXX syndrome. Thus, in many cases, the presence of additional X chromosomes appears to disrupt the delicate balance of genetic information essential to normal female development.

47,XYY Condition

Another human condition involving the sex chromosomes is **47,XYY.** Studies of this condition, in which the only deviation from diploidy is the presence of an additional Y chromosome in an otherwise normal male karyotype, were initiated in 1965 by Patricia Jacobs. She discovered that 9 of 315 males in a Scottish maximum security prison had the 47,XYY karyotype. These males were significantly above average in height and had been incarcerated as a result of dangerous, violent, or criminal propensities. Of the nine males studied, seven were of subnormal intelligence, and all suffered personality disorders. Several other studies produced similar findings.

The possible correlation between this chromosome composition and criminal behavior piqued considerable interest, and extensive investigation of the phenotype and frequency of the 47,XYY condition in both criminal and noncriminal populations ensued. Above-average height (usually over 6 feet) and subnormal intelligence were substantiated, and the frequency of males displaying this karyotype was indeed revealed to be higher in penal and mental institutions compared with unincarcerated populations [one study showed 29 XYY males when 28,366 were examined (0.10%)]. A particularly relevant question involves the characteristics displayed by the XYY males who are not incarcerated. The only nearly constant association is that such individuals are over 6 feet tall.

A study to further address this issue was initiated in 1974 to identify 47,XYY individuals at birth and to follow their behavioral patterns during preadult and adult development. While the study was considered unethical and soon abandoned, it has became clear that there are many XYY males present in the population who do not exhibit antisocial behavior and who lead normal lives. Therefore, we must conclude that there is a high, but not constant, correlation between the extra Y chromosome and the predisposition of these males to exhibit behavioral problems.

Sexual Differentiation in Humans

Once researchers had established that, in humans, it is the Y chromosome that houses genetic information necessary for maleness, they attempted to pinpoint a specific gene or genes capable of providing the “signal” responsible for sex determination. Before we delve into this topic, it is useful to consider how sexual differentiation occurs in order to better comprehend how humans develop into sexually dimorphic males and females. During early development, every human embryo undergoes a period when it is potentially hermaphroditic. By the fifth week of gestation, gonadal primordia (the tissues that will form the gonad) arise as a pair of **gonadal (genital) ridges** associated with each embryonic kidney. The embryo is potentially hermaphroditic because at this stage its gonadal phenotype is sexually indifferent—male or female reproductive structures cannot be distinguished, and the gonadal ridge tissue can develop to form male or female gonads. As development progresses, primordial germ cells migrate to these ridges, where an outer cortex and inner medulla form (*cortex* and *medulla* are the outer and inner tissues of an organ, respectively). The cortex is capable of developing into an ovary, while the medulla may develop into a testis. In addition, two sets of undifferentiated ducts called the Wolffian and Müllerian ducts exist in each embryo. Wolffian ducts differentiate into other organs of the male reproductive tract, while Müllerian ducts differentiate into structures of the female reproductive tract.

Because gonadal ridges can form either ovaries or testes, they are commonly referred to as **bipotential gonads.** What switch triggers gonadal ridge development into testes or ovaries? The presence or absence of a Y chromosome is the key. If cells of the ridge have an XY constitution, development of the medulla into a testis is initiated around the seventh week. However, in the absence of the Y chromosome, no male development occurs, the cortex of the ridge subsequently forms ovarian tissue, and the Müllerian duct forms oviducts (Fallopian tubes), uterus, cervix, and portions of the vagina. Depending on which pathway is initiated, parallel development of the appropriate male or female duct system then occurs, and the other duct system degenerates. If testes differentiation is initiated, the embryonic testicular tissue secretes hormones that are essential for continued male sexual differentiation. As we will discuss in the next section, the presence of a Y chromosome and the development of the testes also inhibit formation of female reproductive organs.

In females, as the twelfth week of fetal development approaches, the oogonia within the ovaries begin meiosis, and primary oocytes can be detected. By the twenty-fifth week of gestation, all oocytes become arrested in meiosis and remain dormant until puberty is reached some 10 to 15 years later. In males, on the other hand, primary spermatocytes are not produced until puberty is reached (see Figure 2.11).

As sexual dimorphism is considered, it is important to distinguish between *primary sexual differentiation,* which involves only the gonads, where gametes are produced, and *secondary sexual differentiation,* which involves the overall phenotype of the organism. Secondary effects include clear differences in such organs as mammary glands and external genitalia as well as other traits that differ between males and females.

The Y Chromosome and Male Development

The human Y chromosome, unlike the X, was long thought to be mostly blank genetically. It is now known that this is not true, even though the Y chromosome contains far fewer genes than does the X. Data from the Human Genome Project indicate that the Y chromosome has at least 75 genes, compared to 900–1400 genes on the X. Current analysis of these genes and regions with potential genetic function reveals that some have homologous counterparts on the X chromosome and others do not. In addition, recent work has revealed that a small number of conserved and essential genes previously thought to be lost from the Y chromosome throughout evolution are present on autosomes. Present on both ends of the Y chromosome are so-called **pseudoautosomal regions (PARs)** that share homology with regions on the X chromosome and synapse and recombine with it during meiosis. The presence of such a pairing region is critical to segregation of the X and Y chromosomes during male gametogenesis. The remainder of the chromosome, about 95 percent of it, does not synapse or recombine with the X chromosome. As a result, it was originally referred to as the *nonrecombining region of the Y (NRY)*. More recently, researchers have designated this region as the **male-specific region of the Y (MSY).** Some portions of the MSY share homology with genes on the X chromosome, and others do not.

The human Y chromosome is diagrammed in Figure 7.3. The MSY is divided about equally between *euchromatic* regions, containing functional genes, and *heterochromatic* regions, lacking genes. Within euchromatin, adjacent to the PAR of the short arm of the Y chromosome, is a critical gene that controls male sexual development, called the **sex-determining region Y (SRY)**. In humans, the absence of a Y chromosome almost always leads to female development; thus, this gene is absent from the X chromosome. At six to eight weeks of development, the *SRY* gene becomes active in XY embryos. *SRY* encodes a protein that causes the undifferentiated gonadal tissue of the embryo to form testes. This protein is called the **testis-determining factor (TDF).** *SRY* (or a closely related version) is present in all mammals thus far examined, indicative of its essential function throughout this diverse group of animals.*

Our ability to identify the presence or absence of DNA sequences in rare individuals whose sex-chromosome composition does not correspond to their sexual phenotype has provided evidence that *SRY* is the gene responsible for male sex determination. For example, there are human males who have two X and no Y chromosomes. Often, attached to one of their X chromosomes is the region of the Y that contains *SRY*. There are also females who have one X and one Y chromosome, a condition known as XY sex reversal or Swyer syndrome. Their Y is almost always missing the *SRY* gene or they have a specific mutation in *SRY*. These observations argue strongly in favor of the role of *SRY* in providing the primary signal for male development.

Further support of this conclusion involves an experiment using **transgenic mice.** These animals are produced from fertilized eggs injected with foreign DNA that is subsequently incorporated into the genetic composition of the developing embryo. In normal mice, a chromosome region designated *Sry* has been identified that is comparable to *SRY* in humans. When mouse DNA containing *Sry* is injected into normal XX mouse eggs, most of the offspring develop into males.

The question of how the product of this gene triggers development of embryonic gonadal tissue into testes rather than ovaries has been under investigation for 25 years. TDF functions as a *transcription factor,* a DNA-binding protein that interacts directly with regulatory sequences of other genes to stimulate their expression. Thus, while TDF behaves as a master switch that controls other genes downstream in the process of sexual differentiation, identifying TDF target genes has been difficult. One potential target for activation by TDF that has been extensively studied is the gene for **Müllerian inhibiting substance (MIS)**, [also called Mullerian inhibiting hormone, (MIH), or anti-Mullerian hormone]. Cells of the developing testes secrete MIS. As its name suggests, MIS protein causes regression (atrophy) of cells in the Müllerian duct. Degeneration of the duct prevents formation of the female reproductive tract.

Other autosomal genes are part of a cascade of genetic expression initiated by *SRY*. Examples include the human *SOX9* gene and the mouse homolog *Sox9*, which when activated by *SRY,* leads to the differentiation of cells that form the seminiferous tubules that contain male germ cells. In the mouse, fibroblast growth factor 9 *(Fgf9)* is upregulated in XY gonads. Testis development is completely blocked in gonads lacking *Fgf9,* and signs of ovarian development occur. Another gene, *SF1,* is involved in the regulation of enzymes affecting steroid metabolism. In mice, this gene is initially active in both the male and female bisexual genital ridge, persisting until the point in development when testis formation is apparent. At that time, its expression persists in males but is extinguished in females. Recent work using mice has

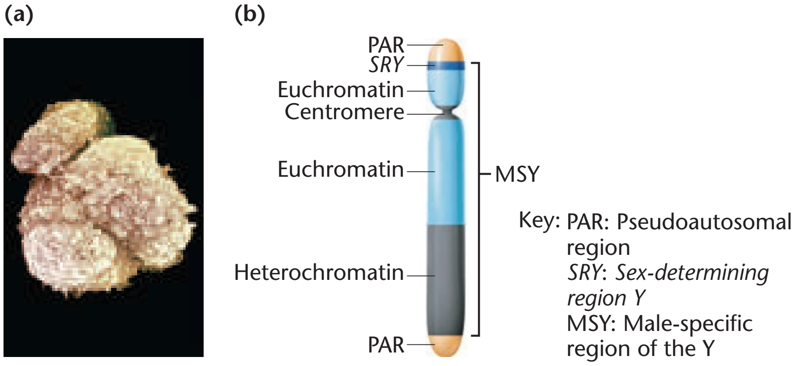

(a)
(b)
PAR
SRY
Euchromatin
Centromere
Euchromatin
MSY
Heterochromatin
PAR
Key: PAR: Pseudoautosomal region
SRY: Sex-determining region Y
MSY: Male-specific region of the Y

FIGURE 7.3 **(a)** Electron micrograph of the human Y chromosome (magnification × 35,000) and **(b)** regions of the Y chromosome.

* It is interesting to note that in chickens, a similar gene has recently been identified. Called *DMRTI*, it is located on the Z chromosome. This gene is the subject of Problem 29 in the Problems section at the end of the chapter.

suggested that testicular development may be actively repressed throughout the life of females by downregulating expression of specific genes. This is based on experiments showing that, in adult female mice, deletion of a gene *Foxl2,* which encodes a transcription factor, leads to transdifferentiation of the ovary into the testis.

In 2016, researchers at the University of Hawaii published novel work demonstrating that two genes in mice, *Sox9* and *Eif2s3y,* could substitute for the Y chromosome. *Sry* activates *Sox9,* and *Eif2s3y* has a homolog on the X chromosome (*Eif2s3x*). Transgenic mice with one X and no Y chromosome were generated. But in these mice, *Sry* was replaced with a transgenic copy of *Sox9* and made to overexpress *Eif2s3x* from an X chromosome, beyond the levels produced normally by the X and Y chromosomes. These males, lacking a Y chromosome, produced haploid male gametes. They did not produce mature sperm but yielded round spermatids that were used to fertilize an oocyte *in vitro,* resulting in viable offspring. This study demonstrated that *Sox9,* in the absence of *Sry,* and the *Eif2s3y* homolog, *Eif2s3x,* allow for male gamete development and initiation of spermatogenesis in the absence of a complete Y chromosome. While these two genes can result in male gametes that produce offspring through assisted reproductive technology, other genes are necessary to produce mature sperm, but nonetheless, experiments such as these are providing novel insights into the genetics of sex-determination pathways. Establishment of the link between these various genes and sex determination has brought us closer to a complete understanding of how males and females arise in humans, but much work remains to be done.

Findings by David Page and his many colleagues have provided a reasonably complete picture of the MSY region of the human Y chromosome. Page has spearheaded the detailed study of the Y chromosome for the past several decades. The MSY consists of about 23 million base pairs (23 Mb) and can be divided into three regions. The first region is the *X-transposed region.* It comprises about 15 percent of the MSY and was originally derived from the X chromosome in the course of human evolution (about 3 to 4 million years ago). The X-transposed region is 99 percent identical to region Xq21 of the modern human X chromosome. Two genes, both with X chromosome homologs, are present in this region.

Research by Page and others has also revealed that sequences called **palindromes**—sequences of base pairs that read the same but in the opposite direction on complementary strands—are present throughout the MSY. Recombination between palindromes on sister chromatids of the Y chromosome during replication is a mechanism used to repair mutations in the Y chromosome. This discovery has fascinating implications concerning how the Y chromosome may maintain its size and structure.

Another interesting finding is that the MSY of the human Y chromosome is very different in sequence structure than the MSY from chimpanzees. The study indicates that rapid evolution has occurred since separation of these species over 6 million years ago—a surprise given that primate sex chromosomes have been in existence for hundreds of millions of years. Over 30 percent of the chimpanzee MSY sequence has no homologous sequence in the human MSY. The chimpanzee MSY has lost many protein-coding genes compared to common ancestors but contains twice the number of palindromic sequences as the human MSY.

The second area of the MSY is designated the *X-degenerative region.* Comprising about 20 percent of the MSY, this region contains DNA sequences that are even more distantly related to those present on the X chromosome. The X-degenerative region contains 27 single-copy genes and a number of *pseudogenes* (genes whose sequences have degenerated sufficiently during evolution to render them nonfunctional). Twenty of the 27 genes located here share homology with counterparts on the X chromosome and evolved from genes on the X chromosome. One of these is the *SRY* gene, discussed earlier. Other X-degenerative genes that encode protein products are expressed ubiquitously in all tissues in the body, but *SRY* is expressed only in the testes.

The third area, the *ampliconic region,* contains about 30 percent of the MSY, including most of the genes closely associated with the development of testes. These genes lack counterparts on the X chromosome, and their expression is limited to the testes. There are 60 transcription units (genes that yield a product) divided among nine gene families in this region, most represented by multiple copies. Members of each family have nearly identical (>98 percent) DNA sequences. Each repeat unit is an **amplicon** and is contained within seven segments scattered across the euchromatic regions of both the short and long arms of the Y chromosome. Genes in the ampliconic region encode proteins specific to the development and function of the testes, and the products of many of these genes are directly related to fertility in males. It is currently believed that a great deal of male sterility in our population can be linked to mutations in these genes.

Until relatively recently it was thought that the Y chromosome only contributed to sex determination and male fertility. A recent area of investigation has involved the Y chromosome and paternal age. For many years, it has been known that maternal age is correlated with an elevated rate of offspring with chromosomal defects, including Down syndrome (see Chapter 8). Advanced paternal age has now been associated with an increased risk in offspring of congenital disorders with a genetic basis, including certain cancers, schizophrenia, autism, and other conditions, collectively known as *paternal age effects (PAE)*. Studies in which the genomes of sperm have been sequenced have demonstrated the presence of specific PAE mutations including numerous ones on the Y chromosome. Evidence suggests that PAE mutations are positively selected for and result in an enrichment of mutant sperm over time.

Similarly, an analysis of blood samples and medical records for more than 6000 men in Sweden revealed a correlation between smoking and complete loss of the Y chromosome in blood cells. Y chromosome loss was also correlated to elevated cancer risk among male smokers, reduced expression of tumor-suppressor genes, and compromised immunity. This and other research provides further evidence that genes on the Y chromosome affect more than sex determination and male fertility.

This recent work has greatly expanded our picture of the genetic information carried by this unique chromosome. It clearly refutes the so-called *wasteland theory,* prevalent some 25 years ago, that depicted the human Y chromosome as almost devoid of genetic information other than a few genes that cause maleness. The knowledge we have gained provides the basis for a much clearer picture of how maleness is determined.

NOW SOLVE THIS

7.1 Campomelic dysplasia (CMD1) is a congenital human syndrome featuring malformation of bone and cartilage. It is caused by an autosomal dominant mutation of a gene located on chromosome 17. Consider the following observations in sequence, and in each case, draw whatever appropriate conclusions are warranted.

(a) Of those with the syndrome who are karyotypically 46,XY, approximately 75 percent are sex reversed, exhibiting a wide range of female characteristics.

(b) The nonmutant form of the gene, called *SOX9*, is expressed in the developing gonad of the XY male, but not the XX female.

(c) The *SOX9* gene shares 71 percent amino acid coding sequence homology with the Y-linked *SRY* gene.

(d) CMD1 patients who exhibit a 46,XX karyotype develop as females, with no gonadal abnormalities.

■ **HINT:** *This problem asks you to apply the information presented in this chapter to a real-life example. The key to its solution is knowing that some genes are activated and produce their normal product as a result of expression of products of other genes found on different chromosomes.*

7.3 The Ratio of Males to Females in Humans Is Not 1.0

The presence of heteromorphic sex chromosomes in one sex of a species but not the other provides a potential mechanism for producing equal proportions of male and female offspring. This potential depends on the segregation of the X and Y (or Z and W) chromosomes during meiosis, such that half of the gametes of the heterogametic sex receive one of the chromosomes and half receive the other one. As we learned in Section 7.2, small pseudoautosomal regions of pairing homology do exist at both ends of the human X and Y chromosomes, suggesting that the X and Y chromosomes do synapse and then segregate into different gametes. Provided that both types of gametes are equally successful in fertilization and that the two sexes are equally viable during development, a 1:1 ratio of male and female offspring should result.

The actual proportion of male to female offspring, referred to as the **sex ratio**, has been assessed in two ways. The **primary sex ratio (PSR)** reflects the proportion of males to females conceived in a population. The **secondary sex ratio** reflects the proportion of each sex that is born. The secondary sex ratio is much easier to determine but has the disadvantage of not accounting for any disproportionate embryonic or fetal mortality.

When the secondary sex ratio in the human population was determined in 1969 by using worldwide census data, it did not equal 1.0. For example, in the Caucasian population in the United States, the secondary ratio was a little less than 1.06, indicating that about 106 males were born for each 100 females. (In 1995, this ratio dropped to slightly less than 1.05.) In the African-American population in the United States, the ratio was 1.025. In other countries, the excess of male births is even greater than is reflected in these values. For example, in Korea, the secondary sex ratio was 1.15.

Despite these ratios, it is possible that the PSR is 1.0 and is altered between conception and birth. For the secondary ratio to exceed 1.0, then, prenatal female mortality would have to be greater than prenatal male mortality. However, when this hypothesis was first examined, it was deemed to be false. In a Carnegie Institute study, reported in 1948, the sex of approximately 6000 embryos and fetuses recovered from miscarriages and abortions was determined, and fetal mortality was actually higher in males. On the basis of the data derived from that study, the PSR in U.S. Caucasians was estimated to be 1.079, suggesting that more males than females are conceived in the human population.

To explain why, researchers considered the assumptions on which the theoretical ratio is based:

1. Because of segregation, males produce equal numbers of X- and Y-bearing sperm.
2. Each type of sperm has equivalent viability and motility in the female reproductive tract.
3. The egg surface is equally receptive to both X- and Y-bearing sperm.

No direct experimental evidence contradicts any of these assumptions.

A PSR favoring male conceptions remained dogma for many decades until, in 2015, a study using an extensive dataset was published determining that the PSR is indeeed 1.0, thus concluding that equal numbers of males and females are conceived. Among other parameters, the examination of the sex of 3-day-old and 6-day-old embryos conceived using assisted reproductive technology provided the most direct assessment. Following conception, however, mortality was then shown to fluctuate between the sexes, until at birth, more males than females are born. Thus, female mortality during embryonic and fetal development exceeds that of males. Clearly, this is a difficult topic to investigate but one of continued interest. For now, the most recent findings are convincing and contradict the earlier studies.

7.4 Dosage Compensation Prevents Excessive Expression of X-Linked Genes in Humans and Other Mammals

The presence of two X chromosomes in normal human females and only one X in normal human males is unique compared with the equal numbers of autosomes present in the cells of both sexes. On theoretical grounds alone, it is possible to speculate that this disparity should create a “genetic dosage” difference between males and females, with attendant problems, for all X-linked genes. There is the potential for females to produce twice as much of each product of all X-linked genes. The additional X chromosomes in both males and females exhibiting the various syndromes discussed earlier in this chapter are thought to compound this dosage problem. Embryonic development depends on proper timing and precisely regulated levels of gene expression. Otherwise, disease phenotypes or embryonic lethality can occur. In this section, we will describe research findings regarding X-linked gene expression that demonstrate a genetic mechanism of **dosage compensation** that balances the dose of X chromosome gene expression in females and males.

Barr Bodies

Murray L. Barr and Ewart G. Bertram’s experiments with cats, as well as Keith Moore and Barr’s subsequent study in humans, demonstrate a genetic mechanism in mammals that compensates for X chromosome dosage disparities. Barr and Bertram observed a darkly staining body in the interphase nerve cells of female cats that was absent in similar cells of males. In humans, this body can be easily demonstrated in female cells derived from the buccal mucosa (cheek cells) or in fibroblasts (undifferentiated connective tissue cells), but not in similar male cells (Figure 7.4). This highly condensed structure, about 1 μm in diameter, lies against the nuclear envelope of interphase cells, and it stains positively for a number of different DNA-binding dyes.

This chromosome structure, called a **sex chromatin body,** or simply a **Barr body,** is an inactivated X chromosome. Susumu Ohno was the first to suggest that the Barr body arises from one of the two X chromosomes. This hypothesis is attractive because it provides a possible mechanism for dosage compensation. If one of the two X chromosomes is inactive in the cells of females, the dosage of genetic information that can be expressed in males and females will be equivalent. Convincing, though indirect, evidence for this hypothesis comes from study of the sex-chromosome syndromes described earlier in this chapter. Regardless of how many X chromosomes a somatic cell possesses, all but one of them appear to be inactivated and can be seen as Barr bodies. For example, no Barr body is seen in the somatic cells of Turner 45,X females; one is seen in Klinefelter 47,XXY males; two in 47,XXX females; three in 48,XXXX females; and so on (Figure 7.5). Therefore, the number of Barr bodies follows an $N - 1$ rule, where N is the total number of X chromosomes present.

Although this apparent inactivation of all but one X chromosome increases our understanding of dosage compensation, it further complicates our perception of other

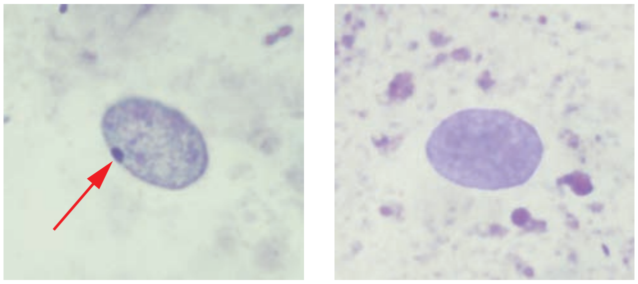


FIGURE 7.4 Photomicrographs comparing cheek epithelial cell nuclei from a male that fails to reveal Barr bodies (right) with a nucleus from a female that demonstrates a Barr body (indicated by the arrow in the left image). This structure, also called a sex chromatin body, represents an inactivated X chromosome.

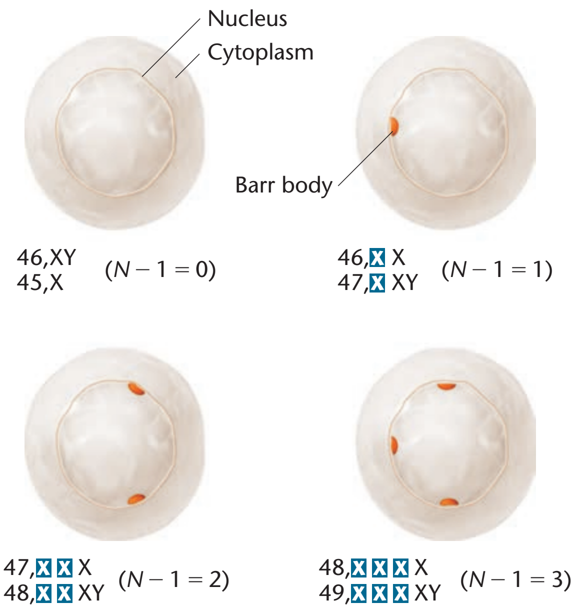


FIGURE 7.5 Occurrence of Barr bodies in various human karyotypes, where all X chromosomes except one (*N* − 1) are inactivated.

matters. For example, because one of the two X chromosomes is inactivated in normal human females, why then is the Turner 45,X individual not entirely normal? Why aren't females with the triplo-X and tetra-X karyotypes (47,XXX and 48,XXXX) completely unaffected by the additional X chromosome? Furthermore, in Klinefelter syndrome (47,XXY), X chromosome inactivation effectively renders the person 46,XY. Why aren't these males unaffected by the extra X chromosome in their nuclei?

One possible explanation is that chromosome inactivation does not normally occur in the very early stages of development of those cells destined to form gonadal tissues. Another possible explanation is that not all genes on each X chromosome forming a Barr body are inactivated. Recent studies have indeed demonstrated that as many as 15 percent of the human X chromosomal genes actually escape inactivation. Clearly, then, not every gene on the X requires inactivation. In either case, excessive expression of certain X-linked genes might still occur at critical times during development despite apparent inactivation of superfluous X chromosomes.

The Lyon Hypothesis

In mammalian females, one X chromosome is of maternal origin, and the other is of paternal origin. Which one is inactivated? Is the inactivation random? Is the same chromosome inactive in all somatic cells? In the early 1960s, Mary Lyon, Liane Russell, and Ernest Beutler independently proposed a hypothesis that answers these questions. They postulated that the inactivation of X chromosomes occurs randomly in somatic cells at a point early in embryonic development, most likely sometime during the blastocyst stage of development. Once inactivation has occurred, all descendant cells have the same X chromosome inactivated as their initial progenitor cell.

This explanation, which has come to be called the **Lyon hypothesis,** was initially based on observations of female mice heterozygous for X-linked coat-color genes. The pigmentation of these heterozygous females was mottled, with large patches expressing the color allele on one X and other patches expressing the allele on the other X. This is the phenotypic pattern that would be expected if different X chromosomes were inactive in adjacent patches of cells. Similar mosaic patterns occur in the black and yellow-orange patches of female tortoiseshell and calico cats (**Figure 7.6**). Such X-linked coat-color patterns do not

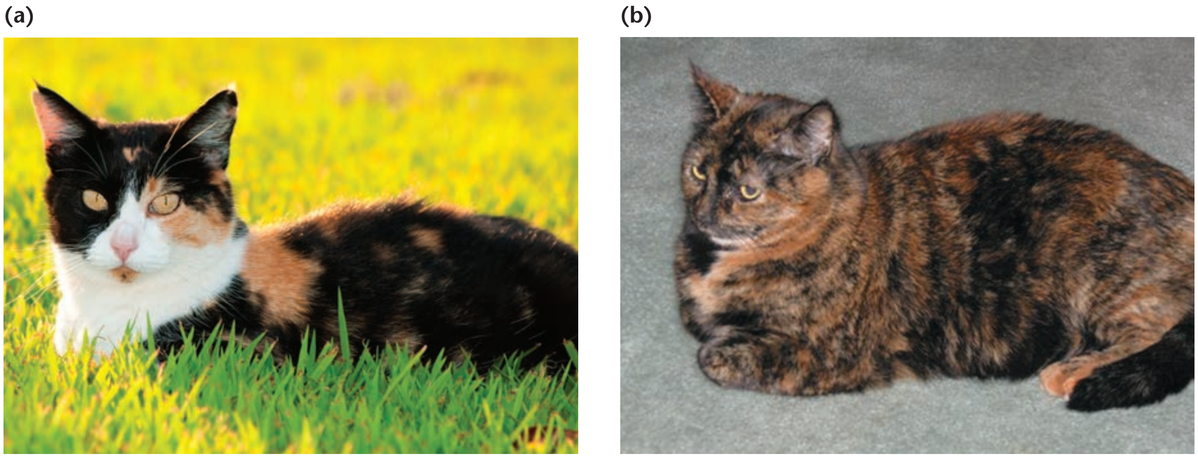


FIGURE 7.6 (a) The random distribution of orange and black patches in a calico cat illustrates the Lyon hypothesis. The white patches are due to another gene, distinguishing calico cats from tortoiseshell cats (b), which lack the white patches.

occur in male cats because all their cells contain the single maternal X chromosome and are therefore hemizygous for only one X-linked coat-color allele.

The most direct evidence in support of the Lyon hypothesis comes from studies of gene expression in clones of human fibroblast cells. Individual cells are isolated following biopsy and cultured *in vitro*. A culture of cells derived from a single cell is called a **clone.** The synthesis of the enzyme glucose-6-phosphate dehydrogenase (G6PD) is controlled by an X-linked gene. Numerous mutant alleles of this gene have been detected, and their gene products can be differentiated from the wild-type enzyme by their migration pattern in an electrophoretic field.

Fibroblasts have been taken from females heterozygous for different allelic forms of *G6PD* and studied. The Lyon hypothesis predicts that if inactivation of an X chromosome occurs randomly early in development, and thereafter all progeny cells have the same X chromosome inactivated as their progenitor, such a female should show two types of clones, each containing only one electrophoretic form of *G6PD,* in approximately equal proportions. This prediction has been confirmed experimentally, and studies involving modern techniques in molecular biology have clearly established that X chromosome inactivation occurs.

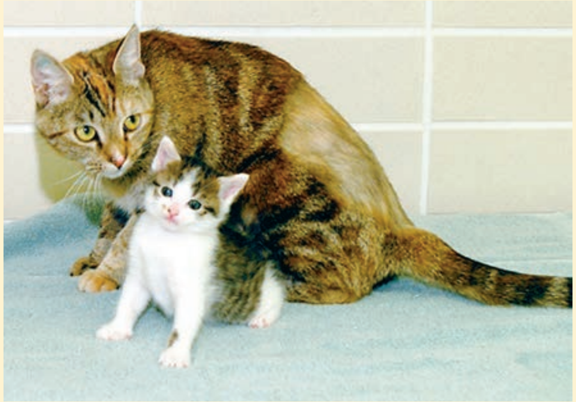
One ramification of X-inactivation is that mammalian females are mosaics for all heterozygous X-linked alleles—some areas of the body express only the maternally derived alleles, and others express only the paternally derived alleles. An especially interesting example involves **red-green color blindness**, an X-linked recessive disorder. In humans, hemizygous males are fully color-blind in all retinal cells. However, heterozygous females display mosaic retinas, with patches of defective color perception and surrounding areas with normal color perception. In this example, random inactivation of one or the other X chromosome early in the development of heterozygous females has led to these phenotypes.

The Mechanism of Inactivation

The least understood aspect of the Lyon hypothesis is the mechanism of X chromosome inactivation. Somehow, either DNA, the attached histone proteins, or both DNA and histone proteins, are chemically modified, silencing most genes that are part of that chromosome. Once silenced, a memory is created that keeps the same homolog inactivated following chromosome replications and cell divisions. Such a process, whereby expression of genes on one homolog, but not the other, is affected, is referred to as **imprinting**. This term also applies to a number of other examples in which genetic information is modified and gene expression is repressed. Collectively, such events are part of the growing field of **epigenetics** (see Chapter 19).

NOW SOLVE THIS

7.2 Carbon Copy (CC), the first cat produced from a clone, was created from an ovarian cell taken from her genetic donor, Rainbow, a calico cat. The diploid nucleus from the cell was extracted and then injected into an enucleated egg. The resulting zygote was then allowed to develop in a petri dish, and the cloned embryo was implanted in the uterus of a surrogate mother cat, who gave birth to CC. CC's surrogate mother was a tabby (see the photo below). Geneticists were very interested in the outcome of cloning a calico cat because they were not certain if the cloned cat would have patches of orange and black, just orange, or just black. Taking into account the Lyon hypothesis, explain the basis of the uncertainty. Would you expect CC to appear identical to Rainbow? Explain why or why not.



Carbon Copy with her surrogate mother.

■ **HINT:** *This problem involves an understanding of the Lyon hypothesis. The key to its solution is to realize that the donor nucleus was from a differentiated ovarian cell of an adult female cat, which itself had inactivated one of its X chromosomes.*

Ongoing investigations are beginning to clarify the mechanism of inactivation. A region of the mammalian X chromosome is the major control unit. This region, located on the proximal end of the p arm in humans, is called the **X-inactivation center (*Xic*)**, and its genetic expression *occurs only on the X chromosome that is inactivated*. The *Xic* is about 1 Mb (10^6 base pairs) in length and is known to contain several putative regulatory units and four genes. One of these, *X-inactive specific transcript (XIST)*, which encodes a long noncoding RNA (lncRNA), is now known to be a critical gene for X-inactivation.

Interesting observations have been made regarding the XIST lncRNA, many coming from experiments that focused on the equivalent gene in the mouse (*Xist*). First, the RNA product is quite large and does not encode a protein, and thus is not translated. The *Xist* transcript is an example of a lncRNA. An *Xist* lncRNA recruits a protein complex that silences transcription at the epigenetic level. The role of lncRNAs in gene expression regulation will be discussed in greater detail in Chapter 19. The *Xist* lncRNAs spread over and coat the X chromosome *bearing the gene that produced them*. Two other noncoding genes at the *Xic* locus, *Tsix* (an antisense partner of *Xist*) and *Xite,* are also believed to play important roles in X-inactivation. It is thought that both *Xist* lncRNA expression and X-linked gene silencing are controlled by a small number of genes that are not inactivated in females.

Another observation is that transcription of *Xist* initially occurs at low levels on all X chromosomes. As the inactivation process begins, however, transcription continues, and is enhanced, only on the X chromosome that becomes inactivated. In addition to silencing genes on the inactivated X chromosome, there are 3D changes in the structure of the inactivated X chromosome that exclude RNA polymerase II from binding to transcription complexes. Recent work has revealed that the *Xist* lncRNA plays a role in these 3D changes in chromosome structure, which in turn help *Xist* to spread and inactivate genes across the X chromosome.

Interesting questions remain regarding imprinting and inactivation. For example, in cells with more than two X chromosomes, what sort of "counting" mechanism exists that designates all but one X chromosome to be inactivated? Studies by Jeannie T. Lee and colleagues suggest that maternal and paternal X chromosomes must first pair briefly and align at their *Xic* loci as a mechanism for counting the number of X chromosomes prior to X-inactivation [**Figure 7.7(a)**]. Using mouse embryonic stem cells, Lee's group deleted the *Tsix* gene contained in the *Xic* locus. This deletion blocked X–X pairing and resulted in chaotic inactivation of 0, 1, or 2 X chromosomes [**Figure 7.7(b)**]. Lee and colleagues provided further evidence for the role of the *Xic* locus in chromosome counting by adding copies of genetically engineered non-X chromosomes containing multiple copies of *Tsix* or *Xite.* (These are referred to as **transgenes** because they are artificially introduced into the organism.) This experimental procedure effectively blocked X–X pairing and prevented X chromosome inactivation [**Figure 7.7(c)**].

Other genes and protein products are being examined for their role in X chromosome pairing and counting. Recent studies by Lee and colleagues have provided evidence that the inactivated X chromosome must associate with regions at the periphery of the nucleus to maintain a state of silenced gene expression. Indeed, in a majority of human female somatic cells the inactivated X chromosome, present as a Barr body, is observed attached to the nuclear envelope.

Many questions remain. What "blocks" the *Xic* locus of the active chromosome, preventing further transcription of *Xist*? How does imprinting impart a memory such that inactivation of the same X chromosome or chromosomes is subsequently maintained in progeny cells, as the Lyon hypothesis calls for? Whatever the answers to these questions, scientists have taken exciting steps toward understanding how dosage compensation is accomplished in mammals.

Finally, modern applications of genomic analysis (which you will learn more about in Chapter 22) have enabled researchers to compare gene expression changes between males and females on a genome-wide scale. Such

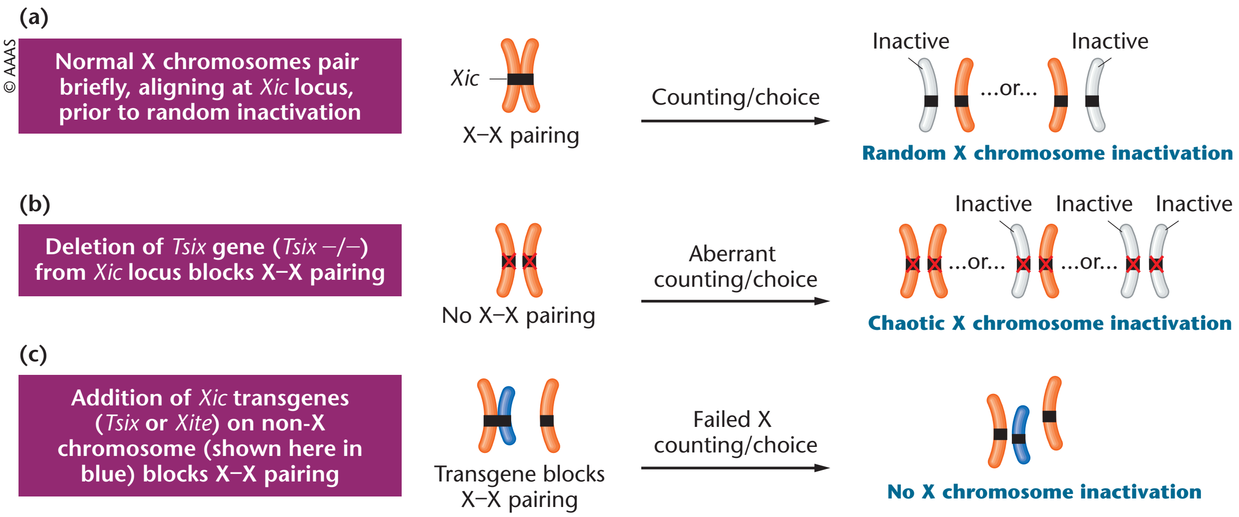

(a)
© AAAS
Normal X chromosomes pair briefly, aligning at *Xic* locus, prior to random inactivation
Xic
X–X pairing
Counting/choice
Inactive
Inactive
...or...
Random X chromosome inactivation
(b)
Deletion of *Tsix* gene (*Tsix* –/–) from *Xic* locus blocks X–X pairing
No X–X pairing
Aberrant counting/choice
Inactive
Inactive
Inactive
...or...
...or...
Chaotic X chromosome inactivation
(c)
Addition of *Xic* transgenes (*Tsix* or *Xite*) on non-X chromosome (shown here in blue) blocks X–X pairing
Transgene blocks X–X pairing
Failed X counting/choice
No X chromosome inactivation

FIGURE 7.7 (a) Transient pairing of X chromosomes may be required for initiating X-inactivation. (b) Deleting the *Tsix* gene of the *Xic* locus prevents X-X pairing and leads to chaotic X-inactivation. (c) Blocking X-X pairing by addition of *Xic*-containing transgenes blocks X-X pairing and prevents X-inactivation.

work is revealing examples of sex-biased gene expression—genes that are expressed predominantly in one sex or another (or at a higher level in one sex). Undoubtedly these studies will provide additional insight about silenced genes and activated genes that contribute to sex determination.

7.5 The Ratio of X Chromosomes to Sets of Autosomes Can Determine Sex

We now discuss two interesting cases where the Y chromosome does not play a role in sex determination. First, in the fruit fly, *Drosophila melanogaster*, even though most males contain a Y chromosome, the Y plays no role. Second, in the roundworm, *Caenorhabditis elegans*, the organism lacks a Y chromosome altogether. In both cases, we shall see that the critical factor is the ratio of X chromosomes to the number of sets of autosomes.

D. melanogaster

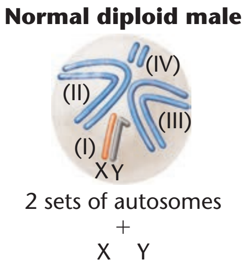
Because males and females in *Drosophila melanogaster* (and other *Drosophila* species) have the same general sex-chromosome composition as humans (males are XY and females are XX), we might assume that the Y chromosome also causes maleness in these flies. However, the elegant work of Calvin Bridges in 1921 showed this not to be true. His studies of flies with quite varied chromosome compositions led him to the conclusion that the Y chromosome is not involved in sex determination in this organism. Instead, Bridges proposed that the X chromosomes and autosomes together play a critical role in sex determination.

Bridges' work can be divided into two phases: (1) A study of offspring resulting from nondisjunction of the X chromosomes during meiosis in females and (2) subsequent work with progeny of females containing three copies of each chromosome, called triploid (3*n*) females. As we have seen previously in this chapter (and as you will see in Figure 8.1), nondisjunction is the failure of paired chromosomes to segregate or separate during the anaphase stage of the first or second meiotic divisions. The result is the production of two types of abnormal gametes, one of which contains an extra chromosome ($n + 1$) and the other of which lacks a chromosome ($n - 1$). Fertilization of such gametes with a haploid gamete produces ($2n + 1$) or ($2n - 1$) zygotes. As in humans, if nondisjunction involves the X chromosome, in addition to the normal complement of autosomes, both an XXY and an X0 sex-chromosome composition may result. (The "0" signifies that neither a second X nor a Y chromosome is present, as occurs in X0 genotypes of individuals with Turner syndrome.)

Contrary to what was later discovered in humans, Bridges found that the XXY flies were normal females and the X0 flies were sterile males. The presence of the Y chromosome in the XXY flies did not cause maleness, and its absence in the X0 flies did not produce femaleness. From these data, Bridges concluded that the Y chromosome in *Drosophila* lacks male-determining factors, but since the X0 males were sterile, it does contain genetic information essential to male fertility.

Bridges was able to clarify the mode of sex determination in *Drosophila* by studying the progeny of triploid females (3*n*), which have three copies each of the haploid complement of chromosomes. *Drosophila* has a haploid number of 4, thereby possessing three pairs of autosomes in addition to its pair of sex chromosomes. Triploid females apparently originate from rare diploid eggs fertilized by normal haploid sperm. Triploid females have heavy-set bodies, coarse bristles, and coarse eyes, and they may be fertile. Because of the odd number of each chromosome (3), during meiosis, a variety of different chromosome complements are distributed into gametes that give rise to offspring with a variety of abnormal chromosome constitutions. Correlations between the sexual morphology and chromosome composition, along with Bridges' interpretation, are shown in **Figure 7.8**.

Bridges realized that the critical factor in determining sex is the ratio of X chromosomes to the number of haploid



Chromosome formulation	Ratio of X chromosomes to autosome sets	Sexual morphology
3X:2A	1.5	Metafemale
3X:3A	1.0	Female
2X:2A	1.0	Female
3X:4A	0.75	Intersex
2X:3A	0.67	Intersex
X:2A	0.50	Male
XY:2A	0.50	Male
XY:3A	0.33	Metamale

FIGURE 7.8 The ratios of X chromosomes to sets of autosomes and the resultant sexual morphology seen in *Drosophila melanogaster.*

sets of autosomes (A) present. Normal (2X:2A) and triploid (3X:3A) females each have a ratio equal to 1.0, and both are fertile. As the ratio exceeds unity (3X:2A, or 1.5, for example), what was once called a *superfemale* is produced. Because such females are most often inviable, they are now more appropriately called **metafemales.**

Normal (XY:2A) and sterile (X0:2A) males each have a ratio of 1:2, or 0.5. When the ratio decreases to 1:3, or 0.33, as in the case of an XY:3A male, infertile **metamales** result. Other flies recovered by Bridges in these studies had an (X:A) ratio intermediate between 0.5 and 1.0. These flies were generally larger, and they exhibited a variety of morphological abnormalities and rudimentary bisexual gonads and genitalia. They were invariably sterile and expressed both male and female morphology, thus being designated as **intersexes.**

Bridges' results indicate that in *Drosophila*, factors that cause a fly to develop into a male are not located on the sex chromosomes but are instead found on the autosomes. Some female-determining factors, however, are located on the X chromosomes. Thus, with respect to primary sex determination, male gametes containing one of each autosome plus a Y chromosome result in male offspring not because of the presence of the Y but because they fail to contribute an X chromosome. This mode of sex determination is explained by the **genic balance theory.** Bridges proposed that a threshold for maleness is reached when the X:A ratio is 1:2 (X:2A), but that the presence of an additional X (XX:2A) alters the balance and results in female differentiation.

Numerous genes involved in sex determination in *Drosophila* have been identified. The recessive autosomal gene *transformer* (*tra*), discovered over 50 years ago by Alfred H. Sturtevant, clearly demonstrated that a single autosomal gene could have a profound impact on sex determination. Females homozygous for *tra* are transformed into sterile males, but homozygous males are unaffected. More recently, another gene, *Sex-lethal* (*Sxl*), has been shown to play a critical role, serving as a "master switch" in sex determination. Activation of the X-linked *Sxl* gene, which relies on a ratio of X chromosomes to sets of autosomes that equals 1.0, is essential to female development. In the absence of activation—as when, for example, the X:A ratio is 0.5—male development occurs.

Although it is not yet exactly clear how this ratio influences the *Sxl* locus, we do have some insights into the question. The *Sxl* locus is part of a hierarchy of gene expression and exerts control over other genes, including *tra* (discussed in the previous paragraph) and *dsx (doublesex)*. The wild-type allele of *tra* is activated by the product of *Sxl* only in females and in turn influences the expression of *dsx*. Depending on how the initial RNA transcript of *dsx* is processed the resultant dsx protein activates either male- or female-specific genes required for sexual differentiation. It is interesting to note that *dsx* homologs have been found in all mammals, thus emphasizing a role for this gene in gonad development.

Each step in this regulatory cascade requires a form of processing called **RNA splicing,** in which portions of the RNA are removed and the remaining fragments are "spliced" back together prior to translation into a protein. In the case of the *Sxl* gene, the RNA transcript may be spliced in different ways, a phenomenon called **alternative splicing.** A different RNA transcript is produced in females than in males. In potential females, the transcript is active and initiates a cascade of regulatory gene expression, ultimately leading to female differentiation. In potential males, the transcript is inactive, leading to a different pattern of gene activity, whereby male differentiation occurs. We will return to this topic in Chapter 18, where alternative splicing is again addressed as one of the mechanisms involved in the regulation of genetic expression in eukaryotes.

Caenorhabditis elegans

The nematode worm *C. elegans* [Figure 7.9(a)] has become a popular organism in genetic studies, particularly for investigating the genetic control of development. Its usefulness is

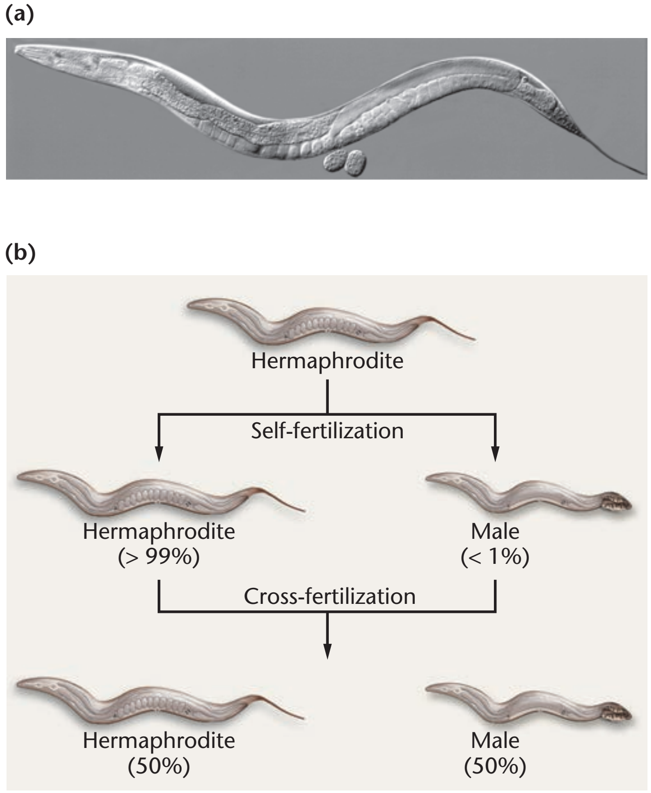


FIGURE 7.9 (a) Photomicrograph of a hermaphroditic nematode, *C. elegans*; (b) the outcomes of self-fertilization in a hermaphrodite, and a mating of a hermaphrodite and a male worm.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Drosophila Sxl Gene Induces Female Development

Donor PGCs	Number of Female Adults with Transplanted Germ-line Cells	Number of Female Adults with Transplanted Germ-line Cells Executing Oogenesis	Number of Female Adults with Transplanted Germ-line Cells Producing Progeny
XY	13	0 (0%)	0
XY + *Sxl*	18	13 (72%)	8
XY-nullo *Sxl*	15	0 (0%)	0
XX	25	25 (100%)	25

Scientists have investigated a critical step involved in determining how germ-cell development, the ability to form either sperm or egg cells, is regulated in fruit flies. As discussed in this chapter, in *Drosophila* the *Sxl* gene encodes a protein that binds to RNA and regulates RNA splicing events. While the molecular mechanism that initiates female germ-cell fate is not known, researchers have hypothesized that the *Sxl* gene product might provide the critical switch that regulates this process. To test this hypothesis, they expressed *Sxl* in **primordial germ cells (PGCs)**, undifferentiated diploid germ cells from XY male flies that would normally become sperm. *Sxl*-containing PGCs were then implanted into the ovaries of female flies and induced to enter oogenesis.

Results:

The table above summarizes *Sxl* implantation results. Implanted cells were tracked to determine whether or not they produced oocytes and whether such oocytes were capable of being fertilized and producing progeny. As shown in the table, XY donor PGCs lacking *Sxl* (XY-nullo *Sxl*) did not produce oocytes when transplanted into ovaries. When XY PGCs containing the *Sxl* gene were transplanted, 72 percent produced eggs, and these eggs were capable of being fertilized and producing progeny! Thus, the addition of a single gene to Y chromosome–containing cells destined to become sperm is sufficient to change the fate of these cells to induce female germ-cell development. Various control experiments were conducted, including XY donor cells with a mutant, nonfunction version of *Sxl* (XY-nullo *Sxl*).

Conclusion:

Sxl functions to initiate development of female germ cells in *Drosophila*. Many questions remain, such as how *Sxl* is activated in female germ cells and which RNA molecules are affected by the *Sxl* gene product in female germ cells. Nonetheless, this result is very exciting because germ-line sexual development is thought to be highly conserved between animal species. Understanding how genes such as *Sxl* regulate germ-cell development in *Drosophila* is expected to help scientists understand key aspects of gamete development in mammals, including humans.

References:

Van Doren, M. (2011). Determining sexual identity. *Science* 333:829–830.
Hashiyama, K., et al. (2011). *Drosophila* sex lethal gene initiates female development in germline progenitors. *Science* 333:885–888

Questions to Consider:

1. Which of the four experiments shown in the above table are controls?
2. Discuss the role of each control experiment. Are the results what you would predict?

based on the fact that adults consist of approximately 1000 cells, the precise lineage of which can be traced back to specific embryonic origins. There are two sexual phenotypes in these worms: males, which have only testes, and hermaphrodites, which contain both testes and ovaries. During larval development of hermaphrodites, testes form that produce sperm, which is stored. Ovaries are also produced, but oogenesis does not occur until the adult stage is reached several days later. The eggs that are produced are fertilized by the stored sperm in a process of self-fertilization.

The outcome of this process is quite interesting [Figure 7.9(b)]. The vast majority of organisms that result are hermaphrodites, like the parental worm; less than 1 percent of the offspring are males. As adults, males can mate with hermaphrodites, producing about half male and half hermaphrodite offspring.

The genetic signal that determines maleness in contrast to hermaphroditic development is provided by genes located on both the X chromosome and autosomes. *C. elegans* lacks a Y chromosome altogether—hermaphrodites have two X chromosomes, while males have only one X chromosome. It is believed that, as in *Drosophila*, it is the ratio of X chromosomes to the number of sets of autosomes that ultimately determines the sex of these worms. A ratio of 1.0 (two X chromosomes and two copies of each autosome) results in hermaphrodites, and a ratio of 0.5 results in males. The absence of a heteromorphic Y chromosome is not uncommon in organisms.

7.6 Temperature Variation Controls Sex Determination in Many Reptiles

We conclude this chapter by discussing several cases involving reptiles, in which the environment—specifically temperature—has a profound influence on sex determination. In contrast to **chromosomal, or genotypic, sex determination (CSD or GSD),** in which sex is determined genetically (as is true of all examples thus far presented in the chapter), the cases that we will now discuss are categorized as **temperature-dependent sex determination (TSD)**. In recent years this topic has received more attention because rapid climate change may influence the sex ratio of certain species and thus threaten their existence in the future.

In many species of reptiles, sex is predetermined at conception by sex-chromosome composition, as is the case in many organisms already considered in this chapter. For example, in many snakes, including vipers, a ZZ/ZW mode is in effect, in which the female is the heterogametic sex (ZW). However, in boas and pythons, it is impossible to distinguish one sex chromosome from the other in either sex. In many lizards, both the XX/XY and ZZ/ZW systems are found, depending on the species.

In still other reptilian species, however, TSD is the norm, including all crocodiles, most turtles, and some lizards, where sex determination is achieved according to the incubation temperature of eggs during a critical period of embryonic development. Three distinct patterns of TSD emerge (cases I–III in **Figure 7.10**). In case I, low temperatures yield 100 percent females, and high temperatures yield 100 percent males. Just the opposite occurs in case II. In case III, low *and* high temperatures yield 100 percent females, while intermediate temperatures yield various proportions of males. The third pattern is seen in various species of crocodiles, turtles, and lizards, although other members of these groups are known to exhibit the other patterns.

Two observations are noteworthy. First, in all three patterns, certain temperatures result in both male and female offspring; second, this pivotal temperature T_P range is fairly narrow, usually spanning less than 5°C, and sometimes only 1°C. The central question raised by these observations is: What are the metabolic or physiological parameters affected by temperature that lead to the differentiation of one sex or the other?

The answer is thought to involve steroids (mainly estrogens) and the enzymes involved in their synthesis. Studies clearly demonstrate that the effects of temperature on estrogens, androgens, and inhibitors of the enzymes controlling their synthesis are involved in the sexual differentiation of ovaries and testes. One enzyme in particular, **aromatase**, converts androgens (male hormones such as testosterone) to estrogens (female hormones such as estradiol). The activity of this enzyme is correlated with the pathway of reactions that occurs during gonadal differentiation activity and is high in developing ovaries and low in developing testes.

In 2016, the first gene linked to TSD, *CIRBP* (cold-inducible RNA-binding protein), was identified in common snapping turtles. Increased expression of this gene occurs within 24 hours when turtle eggs are shifted from a male-determining temperature (MT) to a female-determining temperature (FT). Shortly after increased expression of *CIRBP*, several genes involved in development of the ovary are activated while genes involved in testis development are repressed. Researchers studying *CIRBP*

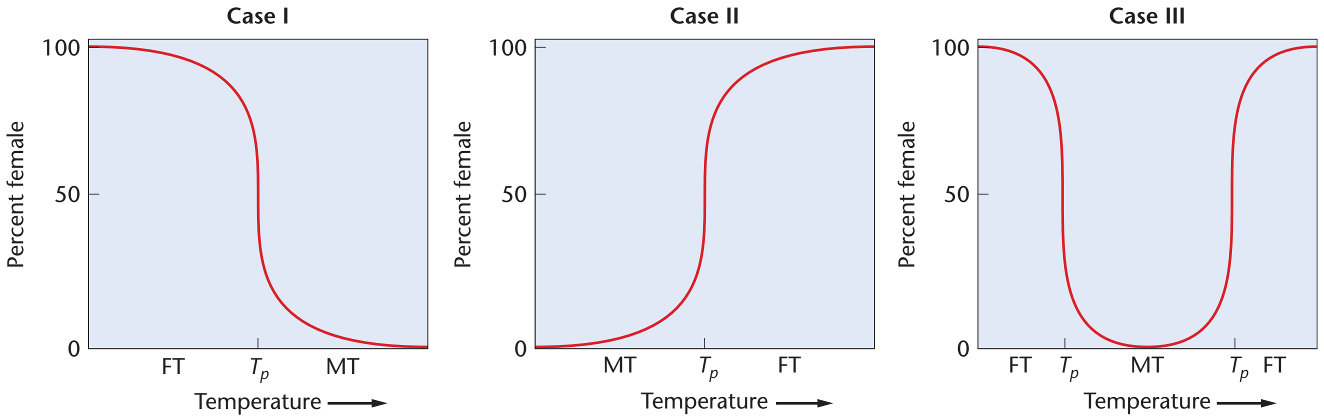


FIGURE 7.10 Three different patterns of temperature-dependent sex determination (TSD) in reptiles, as described in the text. The relative pivotal temperature T_p is crucial to sex determination during a critical point during embryonic development. FT = Female-determining temperature; MT = Male-determining temperature.

also identified an SNP in the gene, which is more common in male than female turtles. Recently, however, researchers discovered that the sex of Australian lizards (*Pogona vitticeps*) can be determined by both sex chromosomes and by the temperature at which *P. vitticeps* eggs are incubated. *P. vitticeps* has a female heterogametic system (ZW) and a male homogametic (ZZ) system. But this research revealed that in the wild nearly 20 percent of ZZ individuals are female and not male. Subsequently it was determined that ZZ females in the wild develop from eggs incubated at elevated temperatures— effectively causing a sex reversal. Hence, this is a species in which sex chromosomes and temperature impact sex determination. Further, experiments show that normal males mated to sex-reversed females produce offspring whose sex is determined solely by temperature. This raises interesting questions about the potential impacts of climate change on sex reversal and sex ratios in this species.

GENETICS, ETHICS, AND SOCIETY

A Question of Gender: Sex Selection in Humans

Throughout history, people have attempted to influence the gender of their unborn offspring by following varied and sometimes bizarre procedures. In medieval Europe, prospective parents would place a hammer under the bed to help them conceive a boy, or a pair of scissors, to conceive a girl. Other practices were based on the ancient belief that semen from the right testicle created male offspring and that from the left testicle created females. In some cultures, efforts to control the sex of offspring has had a darker side—female infanticide. In ancient Greece, the murder of female infants was so common that the male:female ratio in some areas approached 4:1. In some parts of rural India, female infanticide continued up to the 1990s. In 1997, the World Health Organization reported that about 50 million women were "missing" in China, likely because of selective abortion of female fetuses and institutionalized neglect of female children. In recent times, amniocentesis and ultrasound testing, followed by sex-specific abortion, have replaced much of the traditional female infanticide.

New genetic and reproductive technologies offer parents ways to select their children's gender prior to implantation of the embryo in the uterus—methods called *preimplantation gender selection (PGS)*. Following *in vitro* fertilization, embryos can be biopsied and assessed for gender. Only sex-selected embryos are then implanted.

The new PGS methods raise a number of legal and ethical issues. Some people feel that prospective parents have the right to use sex-selection techniques as part of their fundamental procreative liberty. Proponents state that PGS will reduce the suffering of many families. For example, people at risk for transmitting X-linked diseases such as hemophilia or Duchenne muscular dystrophy would be able to enhance their chance of conceiving a female child, who would not express the disease.

The majority of people who undertake PGS, however, do so for nonmedical reasons—to "balance" their families. One argument in favor of this use is that the intentional selection of the sex of an offspring may reduce overpopulation and economic burdens for families who would repeatedly reproduce to get the desired gender. Also, PGS may increase the happiness of both parents and children, as the children would be more "wanted."

On the other hand, some argue that PGS serves neither the individual nor the common good. They argue that PGS is inherently sexist, having its basis in the idea that one sex is superior to the other, and leads to linking a child's worth to gender. Other critics fear that social approval of PGS will open the door to other genetic manipulations of children's characteristics. It is difficult to predict the full effects that PGS will bring to the world. But the gender-selection genie is now out of the bottle and is unwilling to return.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand some of the ethical issues that surround the topic of gender selection.

1. A generally accepted moral and legal concept is that of reproductive autonomy—the freedom to make individual reproductive decisions without external interference. Are there circumstances in which reproductive autonomy should be restricted?

   This question is explored in a series of articles in the American Journal of Bioethics, Vol. 1 (2001). *See the article by* J. A. Robertson on pages 2–9 *for a summary of the moral and legal issues surrounding PGS. Also see* Kalfoglou, A. L., et al. (2013). Ethical arguments for and against sperm sorting for non-medical sex selection: a review. *Repro. BioMed. Online* 26:231–239 (http://www.rbmojournal.com/article/S1472-6483(12)00692-X/fulltext#).

2. If safe and efficient methods of PGS were available, would you use them to help you with family planning? Under what circumstances might you use them?

   A discussion of PGS ethics and methods is presented in Use of reproductive technology for sex selection for nonmedical reasons, Ethics Committee of the American Society for Reproductive Medicine (2015) (http://www.reproductivefacts.org/globalassets/asrm/asrm-content/news-and-publications/ethics-committee-opinions/use_of_reproductive_technology_for_sex_selection_for_nonmedical_reasons-pdfmembers.pdf.).

CASE STUDY Is it a boy or a girl?

Gender is someone's conscious and unconscious feelings of belonging to one sex or another. Each year, about 1 in 4500 children are born with a disorder involving sexual development, where the chromosomal, gonadal, or anatomical sex is atypical. Here we will consider two similar cases with different outcomes. In case 1, a 2-year-old child displayed a mosaic chromosome composition of 45,X/46,XY, with one ovary, one testis, a uterus, and ambiguous genitalia. In case 2, a fetus was diagnosed with a mosaic chromosome composition of 46,XX/47,XXY, and after birth, also displayed one testis, one ovary, a uterus, and ambiguous genitalia. The child in case 1 was adopted from an orphanage and raised as a girl. After consultation with the medical team, the parents decided to continue raising the child as a girl and requested surgery that would completely feminize the child. In case 2, the parents decided to forego treatment and let the child make the choice about gender later in life and to remain neutral about the child's present condition. These cases raise questions about sex determination and the ethics of sex and gender assignment.

1. In humans, what is the role of the MSY region of the Y chromosome in sex determination and gender development?
2. Compare and contrast the ethical decisions faced by the parents in both cases 1 and 2. Should parents be allowed to make the decision about the gender of their child? If not, at what age should the child be allowed to make this decision?

See Kipnis, K., and Diamond, M. (1998). Pediatric ethics and the surgical assignment of sex. *J. Clin. Ethics* 9(4):398–410.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Sexual reproduction depends on the differentiation of male and female structures responsible for the production of male and female gametes, which in turn is controlled by specific genes, most often housed on specific sex chromosomes.
2. Specific sex chromosomes contain genetic information that controls sex determination and sexual differentiation.
3. The presence of a Y chromosome that contains an intact *SRY* gene is responsible for causing maleness in humans and other mammals.
4. In humans, the most current study of the primary sex ratio shows that equal numbers of males and females are conceived, but that more males than females are born.
5. In mammals, female somatic cells randomly inactivate one of two X chromosomes during early embryonic development, a process important for balancing the expression of X chromosome-linked genes in males and females.
6. The Lyon hypothesis states that early in development, inactivation of either the maternal or paternal X chromosome occurs in each cell, and that all progeny cells subsequently inactivate the same chromosome. Mammalian females thus develop as mosaics for the expression of heterozygous X-linked alleles.
7. Although chromosome composition determines the sex of some reptiles, many others show that temperature-dependent effects during egg incubation are critical for sex determination.

INSIGHTS AND SOLUTIONS

1. In *Drosophila,* the X chromosomes may become attached to one another ($\overset{\frown}{XX}$) such that they always segregate together. Some flies thus contain a set of attached X chromosomes plus a Y chromosome.

   (a) What sex would such a fly be? Explain why this is so.

   (b) Given the answer to part (a), predict the sex of the offspring that would occur in a cross between this fly and a normal one of the opposite sex.

   (c) If the offspring described in part (b) are allowed to interbreed, what will be the outcome?

   Solution:

   (a) The fly will be a female. The ratio of X chromosomes to sets of autosomes—which determines sex in *Drosophila*—will be 1.0, leading to normal female development. The Y chromosome has no influence on sex determination in *Drosophila*.

   (b) All progeny flies will have two sets of autosomes along with one of the following sex-chromosome compositions:

   (1) $\overset{\frown}{XX}X$ → a metafemale with 3 X's (called a trisomic)

   (2) $\overset{\frown}{XX}Y$ → a female like her mother

   (3) XY → a normal male

   (4) YY → no development occurs

   (c) A stock will be created that maintains attached-X females generation after generation.

2. The Xg cell-surface antigen is coded for by a gene located on the X chromosome. No equivalent gene exists on the Y chromosome. Two codominant alleles of this gene have been identified: *Xg1* and *Xg2*. A woman of genotype *Xg2/Xg2* bears children with a man of genotype *Xg1/Y*, and they produce a son with Klinefelter syndrome of genotype *Xg1/Xg2/Y*. Using proper genetic terminology, briefly explain how this individual was generated. In which parent and in which meiotic division did the mistake occur?

   Solution: Because the son with Klinefelter syndrome is *Xg1/Xg2/Y*, he must have received both the *Xg1* allele and the Y chromosome from his father. Therefore, nondisjunction must have occurred during meiosis I in the father.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we have focused on sex differentiation, sex chromosomes, and genetic mechanisms involved in sex determination. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, you should answer the following fundamental questions?

   (a) How do we know whether or not a heteromorphic chromosome such as the Y chromosome plays a crucial role in the determination of sex?
   (b) How do we know that in humans the X chromosomes play no role in human sex determination, while the Y chromosome causes maleness and its absence causes femaleness?
   (c) How do we know that *Drosophila* utilizes a different sex-determination mechanism than mammals, even though it has the same sex-chromosome compositions in males and females?
   (d) How do we know that X chromosomal inactivation of either the paternal or maternal homolog is a random event during early development in mammalian females?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p. 151. These all center around sex determination or the expression of genes encoded on sex chromosomes. Write a short essay that discusses sex chromosomes as they contrast with autosomes.

3. Distinguish between the concepts of sexual differentiation and sex determination.

4. Contrast the XX/XY and XX/X0 modes of sex determination.

5. Describe the major difference between sex determination in *Drosophila* and in humans.

6. How do mammals, including humans, solve the “dosage problem” caused by the presence of an X and Y chromosome in one sex and two X chromosomes in the other sex?

7. The phenotype of an early-stage human embryo is considered sexually indifferent. Explain why this is so even though the embryo’s genotypic sex is already fixed.

8. What specific observations (evidence) support the conclusions about sex determination in *Drosophila* and humans?

9. Describe how nondisjunction in human female gametes can give rise to Klinefelter and Turner syndrome offspring following fertilization by a normal male gamete.

10. An insect species is discovered in which the heterogametic sex is unknown. An X-linked recessive mutation for *reduced wing* (*rw*) is discovered. Contrast the F_1 and F_2 generations from a cross between a female with reduced wings and a male with normal-sized wings when

    (a) the female is the heterogametic sex.
    (b) the male is the heterogametic sex.

11. When cows have twin calves of unlike sex (fraternal twins), the female twin is usually sterile and has masculinized reproductive organs. This calf is referred to as a freemartin. In cows, twins may share a common placenta and thus fetal circulation. Predict why a freemartin develops.

12. An attached-X female fly, $\widehat{XX}Y$ (see the “Insights and Solutions” box), expresses the recessive X-linked *white*-eye mutation. It is crossed to a male fly that expresses the X-linked recessive *miniature*-wing mutation. Determine the outcome of this cross in terms of sex, eye color, and wing size of the offspring.

13. Assume that on rare occasions the attached X chromosomes in female gametes become unattached. Based on the parental phenotypes in Problem 12, what outcomes in the F_1 generation would indicate that this has occurred during female meiosis?

14. It has been suggested that any male-determining genes contained on the Y chromosome in humans cannot be located in the limited region that synapses with the X chromosome during meiosis. What might be the outcome if such genes were located in this region?

15. What is a Barr body, and where is it found in a cell?

16. Indicate the expected number of Barr bodies in interphase cells of individuals with Klinefelter syndrome; Turner syndrome; and karyotypes 47,XYY, 47,XXX, and 48,XXXX.

17. Define the Lyon hypothesis.

18. Can the Lyon hypothesis be tested in a human female who is homozygous for one allele of the X-linked *G6PD* gene? Why, or why not?

19. Predict the potential effect of the Lyon hypothesis on the retina of a human female heterozygous for the X-linked red-green color blindness trait.

20. Cat breeders are aware that kittens expressing the X-linked calico coat pattern and tortoiseshell pattern are almost invariably females. Why are they certain of this?

21. In mice, the *Sry* gene (see Section 7.2) is located on the Y chromosome very close to one of the pseudoautosomal regions that pairs with the X chromosome during male meiosis. Given this information, propose a model to explain the generation of unusual males who have two X chromosomes (with an *Sry*-containing piece of the Y chromosome attached to one X chromosome).

22. The genes encoding the red- and green-color-detecting proteins of the human eye are located next to one another on the X chromosome and probably evolved from a common ancestral pigment gene. The two proteins demonstrate 76 percent homology in their amino acid sequences. A normal-visioned woman (with both genes present on each of her two X chromosomes) has a red-color-blind son who was shown to have one copy of the green-detecting gene and no copies of the red-detecting gene. Devise an explanation for these observations at the chromosomal level (involving meiosis).

23. What is the role of the enzyme aromatase in sexual differentiation in reptiles?

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

24. In the wasp *Bracon hebetor*, a form of parthenogenesis (the development of unfertilized eggs into progeny) resulting in haploid organisms is not uncommon. All haploids are males. When offspring arise from fertilization, females almost invariably result. P. W. Whiting has shown that an X-linked gene with nine multiple alleles (X_a, X_b, etc.) controls sex determination. Any homozygous or hemizygous condition results in males, and any heterozygous condition results in females. If an X_a/X_b female mates with an X_a male and lays 50 percent fertilized and 50 percent unfertilized eggs, what proportion of male and female offspring will result?

25. The Amami spiny rat (*Tokudaia osimensis*) lacks a Y chromosome, yet scientists at Hokkaido University in Japan have reported that key sex-determining genes continue to be expressed in this species. Provide possible explanations for why male differentiation can still occur in this mammalian species despite the absence of a Y chromosome.

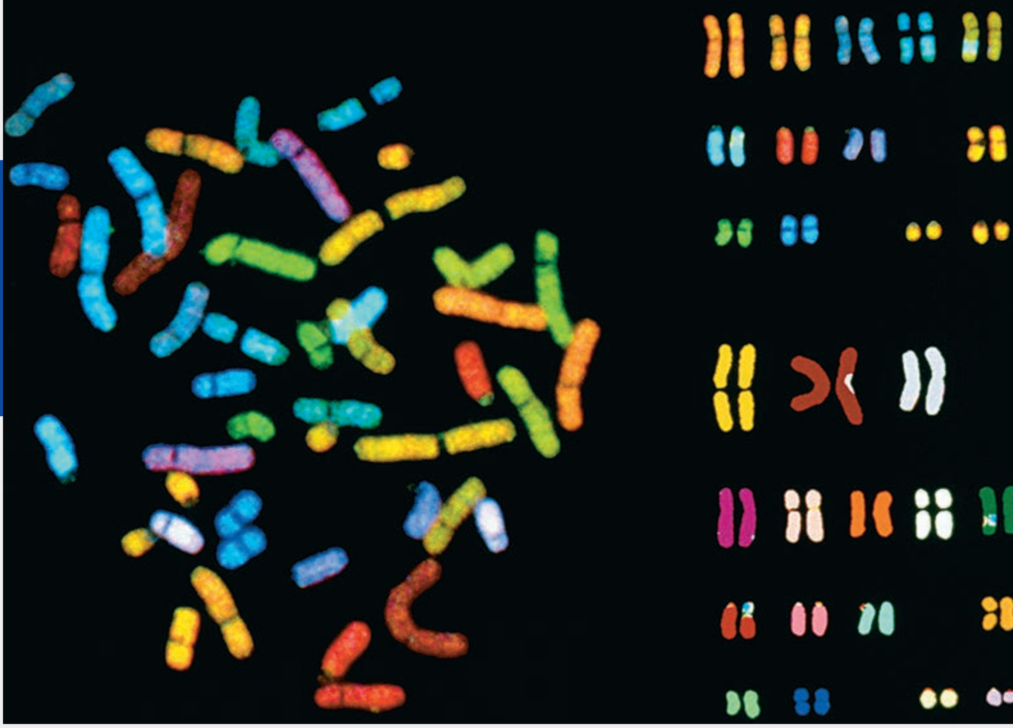
26. In mice, the X-linked dominant mutation *Testicular feminization (Tfm)* eliminates the normal response to the testicular hormone testosterone during sexual differentiation. An XY mouse bearing the *Tfm* allele on the X chromosome develops testes, but no further male differentiation occurs—the external genitalia of such an animal are female. From this information, what might you conclude about the role of the *Tfm* gene product and the X and Y chromosomes in sex determination and sexual differentiation in mammals? Can you devise an experiment, assuming you can "genetically engineer" the chromosomes of mice, to test and confirm your explanation?

27. When the cloned cat Carbon Copy (CC) was born (see the Now Solve This question on p. 161), she had black patches and white patches, but completely lacked any orange patches. The knowledgeable students of genetics were not surprised at this outcome. Starting with the somatic ovarian cell used as the source of the nucleus in the cloning process, explain how this outcome occurred.

28. In reptiles, sex determination was thought to be controlled by sex-chromosome systems or by temperature-dependent sex determination without an inherited component to sex. But as we discussed in section 7.6, in the Australian lizard, *Pogona vitticeps*, it was recently revealed that sex is determined by both chromosome composition and by the temperature at which eggs are incubated. What effects might climate change have on temperature-dependent sex determination in this species, and how might this impact the sex ratio for this species in subsequent generations?

29. In chickens, a key gene involved in sex determination has recently been identified. Called *DMRT1*, it is located on the Z chromosome and is absent on the W chromosome. Like *SRY* in humans, it is male determining. Unlike *SRY* in humans, however, female chickens (ZW) have a single copy while males (ZZ) have two copies of the gene. Nevertheless, it is transcribed only in the developing testis. Working in the laboratory of Andrew Sinclair (a co-discoverer of the human SRY gene), Craig Smith and colleagues were able to "knock down" expression of *DMRT1* in ZZ embryos using RNA interference techniques (see Chapter 18). In such cases, the developing gonads look more like ovaries than testes [*Nature* 461: 267 (2009)]. What conclusions can you draw about the role that the *DMRT1* gene plays in chickens in contrast to the role the *SRY* gene plays in humans?

8



Chromosomal Mutations: Variation in Number and Arrangement

Spectral karyotyping of human chromosomes, utilizing differentially labeled "painting" probes.

CHAPTER CONCEPTS

- The failure of chromosomes to properly separate during meiosis results in variation in the chromosome content of gametes and subsequently in offspring arising from such gametes.
- Plants often tolerate an abnormal genetic content, but, as a result, they often manifest unique phenotypes. Such genetic variation has been an important factor in the evolution of plants.
- In animals, genetic information is in a delicate equilibrium whereby the gain or loss of a chromosome, or part of a chromosome, in an otherwise diploid organism often leads to lethality or to an abnormal phenotype.
- The rearrangement of genetic information within the genome of a diploid organism may be tolerated by that organism but may affect the viability of gametes and the phenotypes of organisms arising from those gametes.
- Chromosomes in humans contain fragile sites—regions susceptible to breakage, which lead to abnormal phenotypes.

In previous chapters, we have emphasized how mutations and the resulting alleles affect an organism's phenotype and how traits are passed from parents to offspring according to Mendelian principles. In this chapter, we look at phenotypic variation that results from more substantial changes than alterations of individual genes—modifications at the level of the chromosome.

Although most members of diploid species normally contain precisely two haploid chromosome sets, many known cases vary from this pattern. Modifications include a change in the total number of chromosomes, the deletion or duplication of genes or segments of a chromosome, and rearrangements of the genetic material either within or among chromosomes. Taken together, such changes are called **chromosome mutations** or **chromosome aberrations,** to distinguish them from gene mutations. Because the chromosome is the unit of genetic transmission, according to Mendelian laws, chromosome aberrations are passed to offspring in a predictable manner, resulting in many unique genetic outcomes.

Because the genetic component of an organism is delicately balanced, even minor alterations of either content or location of genetic information within the genome can result in some form of phenotypic variation. More substantial changes may be lethal, particularly in animals. Throughout this chapter, we consider many types of chromosomal aberrations, the phenotypic consequences for the organism that harbors an aberration, and the impact of the aberration on the offspring of an affected individual. We will also discuss the role of chromosomal aberrations in the evolutionary process.

8.1 Variation in Chromosome Number: Terminology and Origin

Variation in chromosome number ranges from the addition or loss of one or more chromosomes to the addition of one or more haploid sets of chromosomes. Before we embark on our discussion, it is useful to clarify the terminology that describes such changes. In the general condition known as **aneuploidy**, an organism gains or loses one or more chromosomes but not a complete set. The loss of a single chromosome from an otherwise diploid genome is called *monosomy*. The gain of one chromosome results in *trisomy*. These changes are contrasted with the condition of **euploidy**, where complete haploid sets of chromosomes are present. If more than two sets are present, the term **polyploidy** applies. Organisms with three sets are specifically *triploid*, those with four sets are *tetraploid*, and so on. **Table 8.1** provides an organizational framework for you to follow as we discuss each of these categories of aneuploid and euploid variation and the subsets within them.

As we consider cases that include the gain or loss of chromosomes, it is useful to examine how such aberrations originate. For instance, how do the syndromes arise where the number of sex-determining chromosomes in humans is altered (Chapter 7)? As you may recall, the gain (47,XXY) or loss (45,X) of an X chromosome from an otherwise diploid genome affects the phenotype, resulting in **Klinefelter syndrome** or **Turner syndrome,** respectively (see Figure 7.2). Human females may contain extra X chromosomes (e.g., 47,XXX, 48,XXXX), and some males contain an extra Y chromosome (47,XYY).

Such chromosomal variation originates as a random error during the production of gametes, a phenomenon referred to as **nondisjunction**, whereby paired homologs fail to disjoin during segregation. This process disrupts the normal distribution of chromosomes into gametes. The results of nondisjunction during meiosis I and meiosis II for a single chromosome of a diploid organism are shown in **Figure 8.1**. As you can see, abnormal gametes can form containing either two members of the affected chromosome or none at all. Fertilizing these with a normal haploid gamete produces a zygote with either three members (trisomy) or only one member (monosomy) of this chromosome. Nondisjunction leads to a variety of aneuploid conditions in humans and other organisms.

TABLE 8.1 Terminology for Variation in Chromosome Numbers

Term	Explanation
Aneuploidy	$2n \pm x$ chromosomes
Monosomy	$2n - 1$
Disomy	$2n$
Trisomy	$2n + 1$
Tetrasomy, pentasomy, etc.	$2n + 2$, $2n + 3$, etc.
Euploidy	Multiples of n
Diploidy	$2n$
Polyploidy	$3n, 4n, 5n, \ldots$
Triploidy	$3n$
Tetraploidy, pentaploidy, etc.	$4n$, $5n$, etc.
Autopolyploidy	Multiples of the same genome
Allopolyploidy (amphidiploidy)	Multiples of closely related genomes

NOW SOLVE THIS

8.1 A human female with Turner syndrome (47,X) also expresses the X-linked trait hemophilia, as did her father. Which of her parents underwent nondisjunction during meiosis, giving rise to the gamete responsible for the syndrome?

- **HINT:** *This problem involves an understanding of how nondisjunction leads to aneuploidy. The key to its solution is first to review Turner syndrome, discussed above and in more detail later in the text (see Chapter 7), then to factor in that she expresses hemophilia, and finally, to consider which parent contributed a gamete with an X chromosome that underwent normal meiosis.*

For more practice, see Problems 27 and 28.

8.2 Monosomy and Trisomy Result in a Variety of Phenotypic Effects

We turn now to a consideration of variations in the number of autosomes and the genetic consequence of such changes. The most common examples of *aneuploidy,* where an organism has a chromosome number other than an exact multiple of the haploid set, are cases in which a single chromosome is either added to, or lost from, a normal diploid set.

Monosomy

The loss of one chromosome produces a $2n - 1$ complement called **monosomy**. Although monosomy for the X chromosome occurs in humans, as we have seen in 45,X Turner syndrome, monosomy for any of the autosomes is not usually tolerated in humans or other animals. In *Drosophila,* flies that are monosomic for the very small chromosome IV (containing less than 5 percent of the organism's genes) develop more slowly, exhibit reduced body size, and have impaired viability. Monosomy for the larger autosomal chromosomes II and III is apparently lethal because such flies have never been recovered.

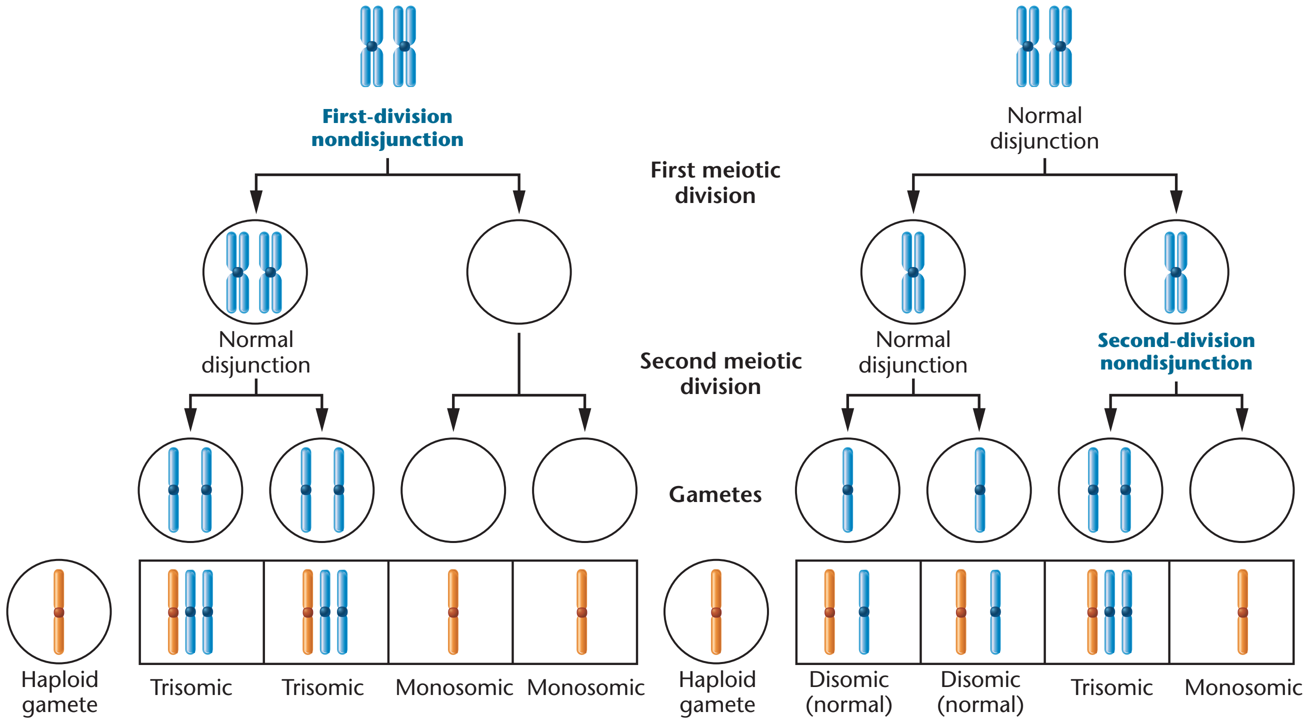


FIGURE 8.1 Nondisjunction during the first and second meiotic divisions. In both cases, some of the gametes that are formed either contain two members of a specific chromosome or lack that chromosome. After fertilization by a gamete with a normal haploid content, monosomic, disomic (normal), or trisomic zygotes are produced.

The failure of monosomic individuals to survive is at first quite puzzling, since at least a single copy of every gene is present in the remaining homolog. However, if just one of those genes is represented by a lethal allele, the unpaired chromosome condition will result in the death of the organism. This will occur because monosomy unmasks recessive lethals that are otherwise tolerated in heterozygotes carrying the corresponding wild-type alleles. Another possible cause of lethality of aneuploidy is that a single copy of a recessive gene may be insufficient to provide adequate function for sustaining the organism, a phenomenon called **haploinsufficiency**.

Aneuploidy is better tolerated in the plant kingdom. Monosomy for autosomal chromosomes has been observed in maize, tobacco, the evening primrose (*Oenothera*), and the jimson weed (*Datura*), among many other plants. Nevertheless, such monosomic plants are usually less viable than their diploid derivatives. Haploid pollen grains, which undergo extensive development before participating in fertilization, are particularly sensitive to the lack of one chromosome and are seldom viable.

Trisomy

In general, the effects of **trisomy** ($2n + 1$) parallel those of monosomy. However, the addition of an extra chromosome produces somewhat more viable individuals in both animal and plant species than does the loss of a chromosome. In animals, this is often true, provided that the chromosome involved is relatively small. However, the addition of a large autosome to the diploid complement in both *Drosophila* and humans has severe effects and is usually lethal during development.

In plants, trisomic individuals are viable, but their phenotype may be altered. A classical example involves the jimson weed, *Datura,* whose diploid number is 24. Twelve primary trisomic conditions are possible, and examples of each one have been recovered. Each trisomy alters the phenotype of the plant's capsule (**Figure 8.2**) sufficiently to produce a unique phenotype. These capsule phenotypes were first thought to be caused by mutations in one or more genes.

Still another example is seen in the rice plant (*Oryza sativa*), which has a haploid number of 12. Trisomic strains for each chromosome have been isolated and studied—the plants of 11 strains can be distinguished from one another and from wild-type plants. Trisomics for the longer chromosomes are the most distinctive, and the plants grow more slowly. This is in keeping with the belief that larger chromosomes cause greater genetic imbalance than smaller ones. Leaf structure, foliage, stems, grain morphology, and plant height also vary among the various trisomies.

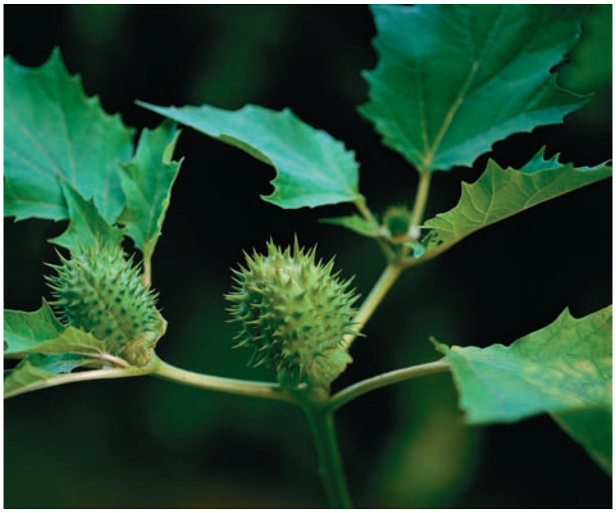


FIGURE 8.2 The capsule of the fruit of the jimson weed, *Datura stramonium,* the phenotype of which is uniquely altered by each of the possible 12 trisomic conditions.

Down Syndrome: Trisomy 21

The only human autosomal trisomy in which a significant number of individuals survive longer than a year past birth was discovered in 1866 by Langdon Down. The condition is now known to result from trisomy of chromosome 21 (Figure 8.3) and is called **Down syndrome** or simply **trisomy 21** (designated 47,21+). This trisomy is found in approximately 1 infant in every 800 live births. While this might seem to be a rare, improbable event, there are approximately 4000–5000 such births annually in the United States, and there are currently over 250,000 individuals with Down syndrome.

Typical of other conditions classified as syndromes, many phenotypic characteristics *may* be present in trisomy 21, but any single affected individual usually exhibits only a subset of these. In the case of Down syndrome, there are 12 to 14 such characteristics, with each individual, on average, expressing 6 to 8 of them. Nevertheless, the outward appearance of these individuals is very similar, and they bear a striking resemblance to one another. This is largely due to a prominent epicanthic fold in each eye* and the typically flat face and round head. Those with Down syndrome are also characteristically short and may have a protruding, furrowed tongue (which causes the mouth to remain partially open) and short, broad hands with characteristic palm and fingerprint patterns. Physical, psychomotor, and mental development is retarded, and poor muscle tone is characteristic. While life expectancy is shortened to an average of about 50 years, individuals are known to survive into their 60s.

Children afflicted with Down syndrome are also prone to respiratory disease and heart malformations, and they show an incidence of leukemia approximately 20 times higher than that of the normal population. In addition, death in older Down syndrome adults is frequently due to Alzheimer disease, the onset of which occurs at a much earlier age than in the normal population.

The Down Syndrome Critical Region (DSCR)

Because Down syndrome is common in our population, a comprehensive understanding of the underlying genetic basis has long been a research goal. Investigations have

*The epicanthic fold, or epicanthus, is a skin fold of the upper eyelid, extending from the nose to the inner side of the eyebrow. It covers and appears to lower the inner corner of the eye, giving the eye a slanted, or almond-shaped, appearance. The epicanthus is a prominent normal component of the eyes in many Asian groups.

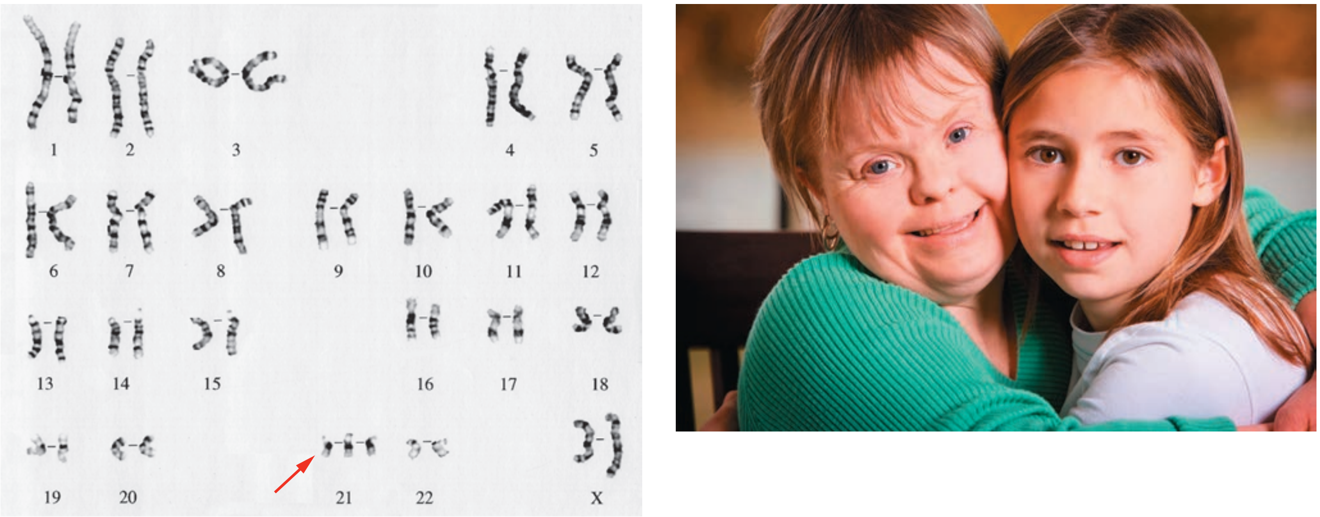


FIGURE 8.3 The karyotype and a photograph of a child with Down syndrome (hugging her unaffected sister on the right). In the karyotype, three members of chromosome 21 are present, creating the 47,21+ condition.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Mouse Models of Down Syndrome

Down syndrome results from the presence of three copies of chromosome 21, leading to the overexpression of some portion of the genes on that chromosome. The resulting gene products are responsible for the multiple characteristics associated with the complete phenotype. Although Down syndrome is not caused by a single gene, individual genes may be responsible for specific characteristics making up the syndrome. Elucidation of which genes are involved has long been the goal of numerous research investigations.

In many of the "Modern Approaches to Understanding Gene Function" features, we focus on experimental approaches that involve the study of one or a few genes. However, geneticists can create model organisms with trisomy, making possible the simultaneous analysis of the function of many genes. Relevant to this discussion, trisomy (Ts) mouse models for Down syndrome have been created and for over two decades have been available to investigators. We focus here on several Ts mouse models that may prove invaluable to our future understanding of the genetic basis of the syndrome and to our developing potential treatments for affected humans.

Human chromosome 21 (Hsa21) and mouse chromosome 16 contain *syntenic* regions—chromosomal regions in different species that contain much of the same content and arrangement of *orthologs,* genes with similar sequences that are present in different species. One of the earliest models created (Ts16) has an extra copy of mouse chromosome 16, and indeed it displays some, but not all, human Down syndrome characteristics. Therefore, Ts16 mice are missing some of the critical orthologs present on Hsa21. Ts16 mice also have extra copies of other genes not present on Hsa21.

Further investigation has established that mouse chromosomes 10 and 17 also contain orthologs for genes on Hsa21. Thus, an important goal of this research effort was to create an extended mouse model that contains most, if not all, of the genes on Hsa21, and in particular, the 33 genes that are part of the Down syndrome critical region (DSCR) discussed earlier in this chapter. This goal was achieved in 2010 when scientists at the Roswell Park Cancer Institute in Buffalo, New York, developed a mouse model of Down syndrome that is trisomic for all syntenic regions of Hsa21 by combining regions of mouse chromosome 10, 16, and 17 (Mmu 10, 16, 17).

Results:

Mmu 10, 16, 17 trisomic mice display the learning and memory deficits characteristic of Down syndrome that were determined through various behavioral tests and electrophysiological recordings of the brain. These mice also show anatomical abnormalities associated with human Down syndrome such as hydrocephalus, accumulation of cerebrospinal fluid in the brain, and a resulting rounded and enlarged cranium.

Conclusion:

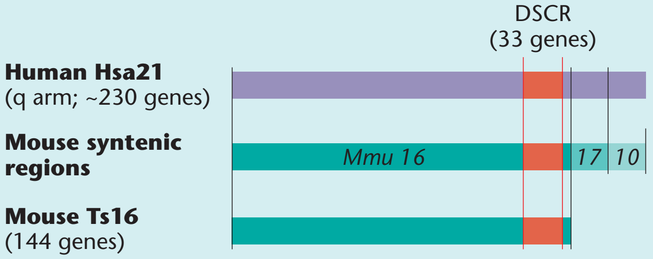
Research with mouse models has demonstrated that no single gene is responsible for all phenotypes associated with Down syndrome, but instead that overexpression of many genes is necessary to produce the full human syndrome. The creation of new models such as the Mmu 10, 16, 17 animals demonstrates their potential value for understanding the genetics of Down syndrome and for developing treatments for Down syndrome patients. Clearly, adding an entire chromosome to an organism's genome has extended our ability to investigate complex human disorders such as Down syndrome.

Reference:

Yu, T. et al. (2010). A mouse model of Down syndrome trisomic for all human chromosome 21 syntenic regions. *Human Mol. Genet.* 19:2780–2791.

Questions to Consider:

1. What are two limitations in creating a mouse model for Down syndrome that is an exact genetic replica of human trisomy 21?
2. How might mouse model animals that are trisomic only for certain syntenic regions of Mmu 10, 16, and 17 be valuable for identifying human orthologs responsible for particular phenotypes observed in Down syndrome?



Schematic representation focusing on the q-arm of human chromosome 21 (Hsa1) where the DSCR is located, compared with syntenic regions of mouse chromosomes (Mmu) 10, 16, and 17, and chromosomal configurations of partial trisomic mouse models.

given rise to the idea that a critical region of chromosome 21 contains the genes that are dosage sensitive in this trisomy and responsible for the many phenotypes associated with the syndrome. This hypothetical portion of the chromosome has been called the **Down syndrome critical region (DSCR).** A mouse model was created in 2004 that is trisomic for the DSCR, although some mice do not exhibit the characteristics of the syndrome. Nevertheless, this remains an important investigative approach, which is explored in the Modern Approaches to Understanding Gene Function box.

Current studies of the DSCR in both humans and mice have led to several interesting findings. Research investigations have now led us to believe that the presence of three copies of the genes present in this region is necessary, but not sufficient for the cognitive deficiencies characteristic of the syndrome. Another finding involves an important observation about Down syndrome—that such individuals have a decreased risk of developing a number of cancers involving solid tumors, including lung cancer and melanoma. This health benefit has been correlated with the presence of an extra copy of the *DSCR1* gene, which encodes a protein that suppresses *vascular endothelial growth factor* (*VEGF*). This suppression, in turn, blocks the process of angiogenesis. As a result, the overexpression of this gene inhibits tumors from forming proper vascularization, diminishing their growth. A 14-year study published in 2002 involving 17,800 individuals with Down syndrome revealed an approximate 10 percent reduction in cancer mortality in contrast to a control population. No doubt, further information will be forthcoming from the study of the DSCR.

The Origin of the Extra Chromosome 21 in Down Syndrome

Most frequently, this trisomic condition occurs through nondisjunction of chromosome 21 during meiosis. Failure of paired homologs to disjoin during either anaphase I or II may lead to gametes with the $n + 1$ chromosome composition. About 75 percent of these errors leading to Down syndrome are attributed to nondisjunction during meiosis I. Subsequent fertilization with a normal gamete creates the trisomic condition.

Chromosome analysis has shown that, while the additional chromosome may be derived from either the mother or father, the ovum is the source in about 95 percent of 47,21+ trisomy cases. Before the development of techniques using polymorphic markers to distinguish paternal from maternal homologs, this conclusion was supported by the more indirect evidence derived from studies of the age of mothers giving birth to infants afflicted with Down syndrome. **Figure 8.4** shows the relationship between the incidence of Down syndrome births and maternal age, illustrating the dramatic increase as the age of the mother increases.

While the frequency is about 1 in 1000 at maternal age 30, a tenfold increase to a frequency of 1 in 100 is noted at age 40. The frequency increases still further to about 1 in 30 at age 45. A very alarming statistic is that as the age of childbearing women exceeds 45, the probability of a Down syndrome birth continues to increase substantially. In spite of these statistics, substantially more than half of Down syndrome births occur to women younger than 35 years, because the overwhelming proportion of pregnancies in the general population involve women under that age.

Although the nondisjunctional event clearly increases with age, we do not know with certainty why this is so. However, one observation is clearly relevant. Meiosis is initiated in all the eggs of a human female when she is still a fetus, until the point where the homologs synapse and recombination has begun. Then oocyte development is arrested in meiosis I. Thus, all primary oocytes have been formed by birth. When ovulation begins at puberty, meiosis is reinitiated in one egg during each ovulatory cycle and continues into meiosis II. The process is once again arrested after ovulation and is not completed unless fertilization occurs.

The end result of this progression is that each ovum that is released has been arrested in meiosis I for about a month longer than the one released during the preceding cycle. As a consequence, women 30 or 40 years old produce ova that are significantly older and that have been

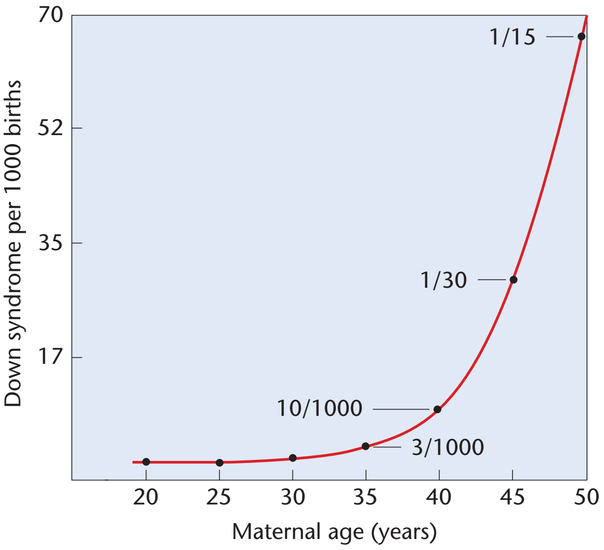


FIGURE 8.4 Incidence of Down syndrome births related to maternal age.

arrested longer than those they ovulated 10 or 20 years previously. In spite of the logic underlying this hypothesis explaining the cause of the increased incidence of Down syndrome as women age, it remains difficult to prove directly.

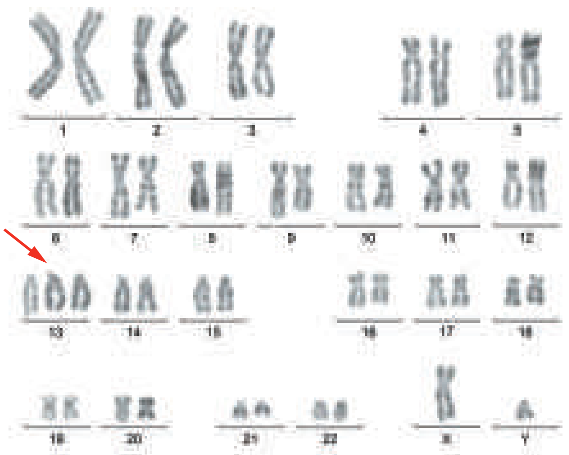
These statistics obviously pose a serious problem for the woman who becomes pregnant late in her reproductive years. Genetic counseling early in such pregnancies is highly recommended. Counseling informs prospective parents about the probability that their child will be affected and educates them about Down syndrome. Although some individuals with Down syndrome must be institutionalized, others benefit greatly from special education programs and may be cared for at home. (Down syndrome children in general are noted for their affectionate, loving nature.) A genetic counselor may also recommend a prenatal diagnostic technique in which fetal cells are isolated and cultured.

In **amniocentesis** and **chorionic villus sampling (CVS)**, the two most familiar approaches, fetal cells are obtained from the amniotic fluid or the chorion of the placenta, respectively. In a newer approach, fetal cells and DNA are derived directly from the maternal circulation, a technique referred to as **noninvasive prenatal genetic diagnosis (NIPGD)**. Requiring only a 10-mL maternal blood sample, this procedure will become increasingly more common because it poses no risk to the fetus. After fetal cells are obtained and cultured, the karyotype can be determined by cytogenetic analysis. If the fetus is diagnosed as being affected, a therapeutic abortion is one option currently available to parents. Obviously, this is a difficult decision involving a number of religious and ethical issues.

Since Down syndrome is caused by a random error—nondisjunction of chromosome 21 during maternal or paternal meiosis—the occurrence of the disorder is *not* expected to be inherited. Nevertheless, Down syndrome occasionally runs in families. These instances, referred to as *familial Down syndrome,* involve a translocation of chromosome 21, another type of chromosomal aberration, which we will discuss later in the chapter.

Human Aneuploidy

Besides Down syndrome, only two human trisomies, and no monosomies, survive to term: **Patau** and **Edwards syndromes** (47,13+ and 47,18+, respectively). Even so, these individuals manifest severe malformations and early lethality. Figure 8.5 illustrates the abnormal karyotype and the many defects characterizing Patau infants.



Mental retardation
Growth failure
Low-set, deformed ears
Deafness
Atrial septal defect
Ventricular septal defect
Abnormal polymorphonuclear granulocytes

Microcephaly
Cleft lip and palate
Polydactyly
Deformed finger nails
Kidney cysts
Double ureter
Umbilical hernia
Developmental uterine abnormalities
Cryptorchidism

FIGURE 8.5 The karyotype and phenotypic description of an infant with Patau syndrome, where three members of chromosome 13 are present, creating the 47,13+ condition.

The above observation leads us to ask whether many other aneuploid conditions arise but that the affected fetuses do not survive to term. That this is the case has been confirmed by karyotypic analysis of spontaneously aborted fetuses. These studies reveal two striking statistics: (1) Approximately 20 percent of all conceptions terminate in spontaneous abortion (some estimates are considerably higher); and (2) about 30 percent of all spontaneously aborted fetuses demonstrate some form of chromosomal imbalance. This suggests that at least 6 percent (0.20 × 0.30) of conceptions contain an abnormal chromosome complement. A large percentage of fetuses demonstrating chromosomal abnormalities are aneuploids.

An extensive review of this subject by David H. Carr has revealed that a significant percentage of aborted fetuses are trisomic for one of the chromosome groups. Trisomies for every human chromosome have been recovered. Interestingly, the monosomy with the highest incidence among abortuses is the 45,X condition, which produces an infant with Turner syndrome, if the fetus survives to term. Autosomal monosomies are seldom found, however, even though nondisjunction should produce $n - 1$ gametes with a frequency equal to $n + 1$ gametes. This finding suggests that gametes lacking a single chromosome are functionally impaired to a serious degree or that the embryo dies so early in its development that recovery occurs infrequently. We discussed the potential causes of monosomic lethality earlier in this chapter. Carr's study also found various forms of polyploidy and other miscellaneous chromosomal anomalies.

These observations support the hypothesis that normal embryonic development requires a precise diploid complement of chromosomes to maintain the delicate equilibrium in the expression of genetic information. The prenatal mortality of most aneuploids provides a barrier against the introduction of these genetic anomalies into the human population.

8.3 Polyploidy, in Which More Than Two Haploid Sets of Chromosomes Are Present, Is Prevalent in Plants

The term *polyploidy* describes instances in which more than two multiples of the haploid chromosome set are found. The naming of polyploids is based on the number of sets of chromosomes found: A triploid has 3*n* chromosomes; a tetraploid has 4*n;* a pentaploid, 5*n;* and so forth (Table 8.1). Several general statements can be made about polyploidy. This condition is relatively infrequent in many animal species but is well known in lizards, amphibians, and fish, and is much more common in plant species. Usually, odd numbers of chromosome sets are not reliably maintained from generation to generation because a polyploid organism with an uneven number of homologs often does not produce genetically balanced gametes. For this reason, triploids, pentaploids, and so on, are not usually found in plant species that depend solely on sexual reproduction for propagation.

Polyploidy originates in two ways: (1) the addition of one or more extra sets of chromosomes, identical to the normal haploid complement of the same species, resulting in **autopolyploidy**; or (2) the combination of chromosome sets from different species occurring as a consequence of hybridization, resulting in **allopolyploidy** (from the Greek word *allo,* meaning "other" or "different"). The distinction between auto- and allopolyploidy is based on the genetic origin of the extra chromosome sets, as shown in Figure 8.6.

In our discussion of polyploidy, we use the following symbols to clarify the origin of additional chromosome sets. For example, if *A* represents the haploid set of chromosomes of any organism, then

$$A = a_1 + a_2 + a_3 + a_4 + \cdots + a_n$$

where a_1, a_2, and so on, are individual chromosomes and *n* is the haploid number. A normal diploid organism is represented simply as *AA*.

Autopolyploidy

In autopolyploidy, each additional set of chromosomes is identical to the parent species. Therefore, triploids are represented as *AAA,* tetraploids are *AAAA,* and so forth.

Autotriploids arise in several ways. A failure of all chromosomes to segregate during meiotic divisions can produce a diploid gamete. If such a gamete is fertilized by a haploid gamete, a zygote with three sets of chromosomes is produced. Or, rarely, two sperm may fertilize an ovum, resulting in a triploid zygote. Triploids are also produced under experimental conditions by crossing diploids with tetraploids. Diploid organisms produce

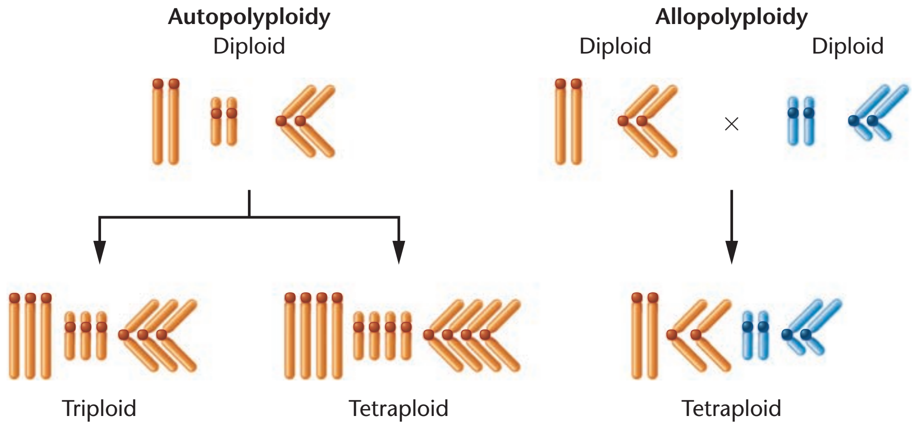


FIGURE 8.6 Contrasting chromosome origins of an autopolyploid versus an allopolyploid karyotype.

gametes with n chromosomes, while tetraploids produce $2n$ gametes. Upon fertilization, the desired triploid is produced.

Because they have an even number of chromosomes, **autotetraploids** ($4n$) are theoretically more likely to be found in nature than are autotriploids. Unlike triploids, which often produce genetically unbalanced gametes with odd numbers of chromosomes, tetraploids are more likely to produce balanced gametes when involved in sexual reproduction.

How polyploidy arises naturally is of great interest to geneticists. In theory, if chromosomes have replicated, but the parent cell never divides and instead reenters interphase, the chromosome number will be doubled. That this very likely occurs is supported by the observation that tetraploid cells can be produced experimentally from diploid cells. This process is accomplished by applying cold or heat shock to meiotic cells or by applying colchicine to somatic cells undergoing mitosis. **Colchicine**, an alkaloid derived from the autumn crocus, interferes with spindle formation, and thus replicated chromosomes cannot separate at anaphase and do not migrate to the poles. When colchicine is removed, the cell can reenter interphase. When the paired sister chromatids separate and uncoil, the nucleus contains twice the diploid number of chromosomes and is therefore $4n$. This process is shown in **Figure 8.7**.

In general, autopolyploids are larger than their diploid relatives. This increase seems to be due to larger cell size rather than greater cell number. Although autopolyploids do not contain new or unique information compared with their diploid relatives, the flower and fruit of plants are often increased in size, making such varieties of greater horticultural or commercial value. Economically important triploid plants include several potato species of the genus *Solanum,* Winesap apples, commercial bananas, seedless watermelons, and the cultivated tiger lily *Lilium tigrinum.* These plants are propagated asexually. Diploid bananas contain hard seeds, but the commercial triploid "seedless" variety has edible seeds. Tetraploid alfalfa, coffee, peanuts, and McIntosh apples are also of economic value because they are either larger or grow more vigorously than do their diploid or triploid counterparts. Many of the most popular varieties of hosta plant are tetraploid. In each case, leaves are thicker and larger, the foliage is more vivid, and the plant grows more vigorously. The commercial strawberry is an octoploid.

We have long been curious about how cells with increased ploidy values, where no new genes are present, express different phenotypes from their diploid counterparts. Our current ability to examine gene expression using modern biotechnology has provided some interesting insights. For example, Gerald Fink and his colleagues have been able to create strains of the yeast *Saccharomyces cerevisiae* with one, two, three, or four copies of the genome. Thus, each strain contains identical genes (they are said to be isogenic) but different ploidy values. These scientists then examined the expression levels of all genes during the entire cell cycle of the organism. Using the rather stringent standards of a tenfold increase or decrease of gene expression, Fink and coworkers proceeded to identify ten cases where, as ploidy increased, gene expression was increased at least tenfold and seven cases where it was reduced by a similar level.

Among these cases are two genes that encode **G1 cyclins**, which are repressed as ploidy increases. These proteins facilitate the cell's movement through the cell cycle, which is thus delayed. The polyploid cell stays in the G1 stage longer than normal, and the cell grows to a larger size. Yeast cells also show different morphology as ploidy increases. Several of the other genes, repressed as ploidy increases, have been linked to cytoskeletal dynamics that account for the morphological changes.

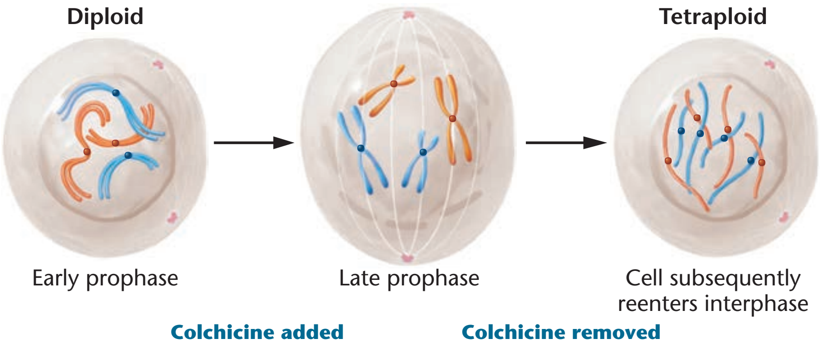


FIGURE 8.7 The potential involvement of colchicine in doubling the chromosome number. Two pairs of homologous chromosomes are shown. While each chromosome had replicated its DNA earlier during interphase, the chromosomes do not appear as double structures until late prophase. When anaphase fails to occur normally, the chromosome number doubles if the cell reenters interphase.

Allopolyploidy

Polyploidy can also result from hybridizing two closely related species. If a haploid ovum from a species with chromosome sets AA is fertilized by sperm from a species with sets BB, the resulting hybrid is AB, where $A = a_1, a_2, a_3, \ldots, a_n$ and $B = b_1, b_2, b_3, \ldots, b_n$. The hybrid plant may be sterile because of its inability to produce viable gametes. Most often, this occurs when some or all of the *a* and *b* chromosomes are not homologous and therefore cannot synapse in meiosis. As a result, unbalanced genetic conditions result. If, however, the new *AB* genetic combination undergoes a natural or an induced chromosomal doubling, two copies of all *a* chromosomes and two copies of all *b* chromosomes will be present, and they will pair during meiosis. As a result, a fertile *AABB* tetraploid is produced. These events are shown in **Figure 8.8**. Since this polyploid contains the equivalent of four haploid genomes derived from separate species, such an organism is called an **allotetraploid**. When both original species are known, an equivalent term, **amphidiploid**, is preferred in describing the allotetraploid.

Amphidiploid plants are often found in nature. Their reproductive success is based on their potential for forming balanced gametes. Since two homologs of each specific chromosome are present, meiosis occurs normally (Figure 8.8) and fertilization successfully propagates the plant sexually. This discussion assumes the simplest situation, where none of the chromosomes in set *A* are homologous to those in set *B*. In amphidiploids, formed from closely related species, some homology between *a* and *b* chromosomes is likely. Allopolyploids are rare in most animals because mating behavior is most often species-specific, and thus the initial step in hybridization is unlikely to occur.

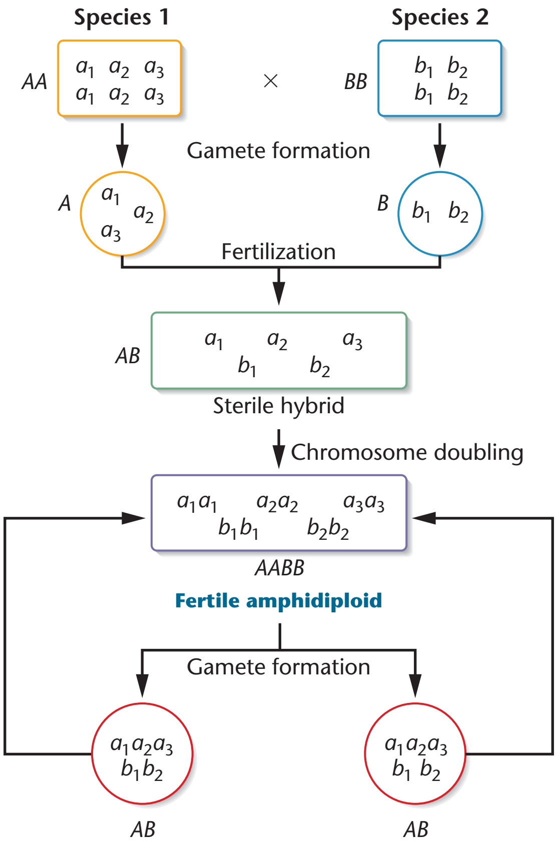


FIGURE 8.8 The origin and propagation of an amphidiploid. Species 1 contains genome *A* consisting of three distinct chromosomes, a_1,a_2, and a_3. Species 2 contains genome *B* consisting of two distinct chromosomes, b_1 and b_2. Following fertilization between members of the two species and chromosome doubling, a fertile amphidiploid containing two complete diploid genomes (*AABB*) is formed.

A classical example of amphidiploidy in plants is the cultivated species of American cotton, *Gossypium* (**Figure 8.9**). This species has 26 pairs of chromosomes: 13 are large and 13 are much smaller. When it was discovered that Old World cotton had only 13 pairs of large chromosomes, allopolyploidy was suspected. After an examination of wild American cotton revealed 13 pairs of small chromosomes, this speculation was strengthened. J. O. Beasley reconstructed the origin of cultivated cotton experimentally by crossing the Old World strain with the wild American strain and then treating the hybrid with colchicine to double the chromosome number. The result of these treatments was a fertile amphidiploid variety of cotton. It contained 26 pairs of chromosomes as well as characteristics similar to the cultivated variety.

Amphidiploids often exhibit traits of both parental species. An interesting example, but one with no practical economic importance, is that of the hybrid formed between the radish *Raphanus sativus* and the cabbage *Brassica oleracea*. Both species have a haploid number $n = 9$. The initial hybrid consists of nine *Raphanus* and nine *Brassica* chromosomes (9R + 9B). Although hybrids are almost always sterile, some fertile amphidiploids (18R + 18B) have been produced. Unfortunately, the root of this plant is more like the cabbage and its shoot is more like the radish; had the converse occurred, the hybrid might have been of economic importance.

FIGURE 8.9 The pods of the amphidiploid form of *Gossypium,* the cultivated American cotton plant.

A much more successful commercial hybridization uses the grasses wheat and rye. Wheat (genus *Triticum*) has a basic haploid genome of seven chromosomes. In addition to normal diploids ($2n = 14$), cultivated autopolyploids exist, including tetraploid ($4n = 28$) and hexaploid ($6n = 42$) species. Rye (genus *Secale*) also has a genome consisting of seven chromosomes. The only cultivated species is the diploid plant ($2n = 14$).

Using the technique outlined in Figure 8.8, geneticists have produced various hybrids. When tetraploid wheat is crossed with diploid rye and the F_1 plants are treated with colchicine, a hexaploid variety ($6n = 42$) is obtained; the hybrid, designated *Triticale,* represents a new genus. Fertile hybrid varieties derived from various wheat and rye species can be crossed or backcrossed. These crosses have created many variations of the genus *Triticale*. The hybrid plants demonstrate characteristics of both wheat and rye. For example, certain hybrids combine the high-protein content of wheat with rye's high content of the amino acid lysine. (The lysine content is low in wheat and thus is a limiting nutritional factor.) Wheat is considered to be a high-yielding grain, whereas rye is noted for its versatility of growth in unfavorable environments. *Triticale* species, which combine both traits, have the potential of significantly increasing grain production. Programs designed to improve crops through hybridization have long been under way in several developing countries.

Endopolyploidy

Endopolyploidy is the condition in which only certain cells in an otherwise diploid organism are polyploid. In such cells, the set of chromosomes replicates repeatedly without nuclear division. Numerous examples of naturally occurring endopolyploidy have been observed. For example, vertebrate liver cell nuclei, including those in humans, often contain $4n$, $8n$, or $16n$ chromosome sets. The stem and parenchymal tissue of apical regions of flowering plants are also often endopolyploid. Cells lining the gut of mosquito larvae attain a $16n$ ploidy, but during the pupal stages, such cells undergo very quick reduction divisions, giving rise to smaller diploid cells. In the water strider *Gerris,* wide variations in chromosome numbers are found in different tissues, with as many as 1024 to 2048 copies of each chromosome in the salivary gland cells. Since the diploid number in this organism is 22, the nuclei of these cells may contain more than 40,000 chromosomes.

Although the role of endopolyploidy is not clear, the proliferation of chromosome copies often occurs in cells where high levels of certain gene products are required.

NOW SOLVE THIS

8.2 When two plants belonging to the same genus but different species are crossed, the F_1 hybrid is more viable and has more ornate flowers. Unfortunately, this hybrid is sterile and can only be propagated by vegetative cuttings. Explain the sterility of the hybrid and what would have to occur for the sterility of this hybrid to be reversed.

- **HINT:** *This problem involves an understanding of allopolyploid plants. The key to its solution is to focus on the origin and composition of the chromosomes in the F_1 and how they might be manipulated.*

8.4 Variation Occurs in the Composition and Arrangement of Chromosomes

The second general class of chromosomal aberrations includes changes that delete, add, or rearrange substantial portions of one or more chromosomes. Included in this broad category are deletions and duplications of genes or part of a chromosome and rearrangements of genetic material in which a chromosome segment is inverted, exchanged with a segment of a nonhomologous chromosome, or merely transferred to another chromosome. Exchanges and transfers are called *translocations*, in which the locations of genes are altered within the genome. These types of chromosome alterations are illustrated in **Figure 8.10**.

In most instances, these structural changes are due to one or more breaks along the axis of a chromosome, followed by either the loss or rearrangement of genetic material. Chromosomes can break spontaneously, but the rate of breakage may increase in cells exposed to chemicals or radiation. Although the actual ends of chromosomes, known as telomeres, do not readily fuse with newly created ends of "broken" chromosomes or with other telomeres, the ends produced at points of breakage are "sticky" and can rejoin other broken ends. If breakage and rejoining do not reestablish the original relationship and if the alteration occurs in germ cells, the gametes will contain the structural rearrangement, which is heritable.

If the aberration is found in one homolog, but not the other, the individual is said to be *heterozygous for the aberration.* In such cases, unusual but characteristic pairing configurations are formed during meiotic synapsis. These patterns are useful in identifying the type of change that has occurred. If no loss or gain of genetic material occurs, individuals bearing the aberration "heterozygously" are likely

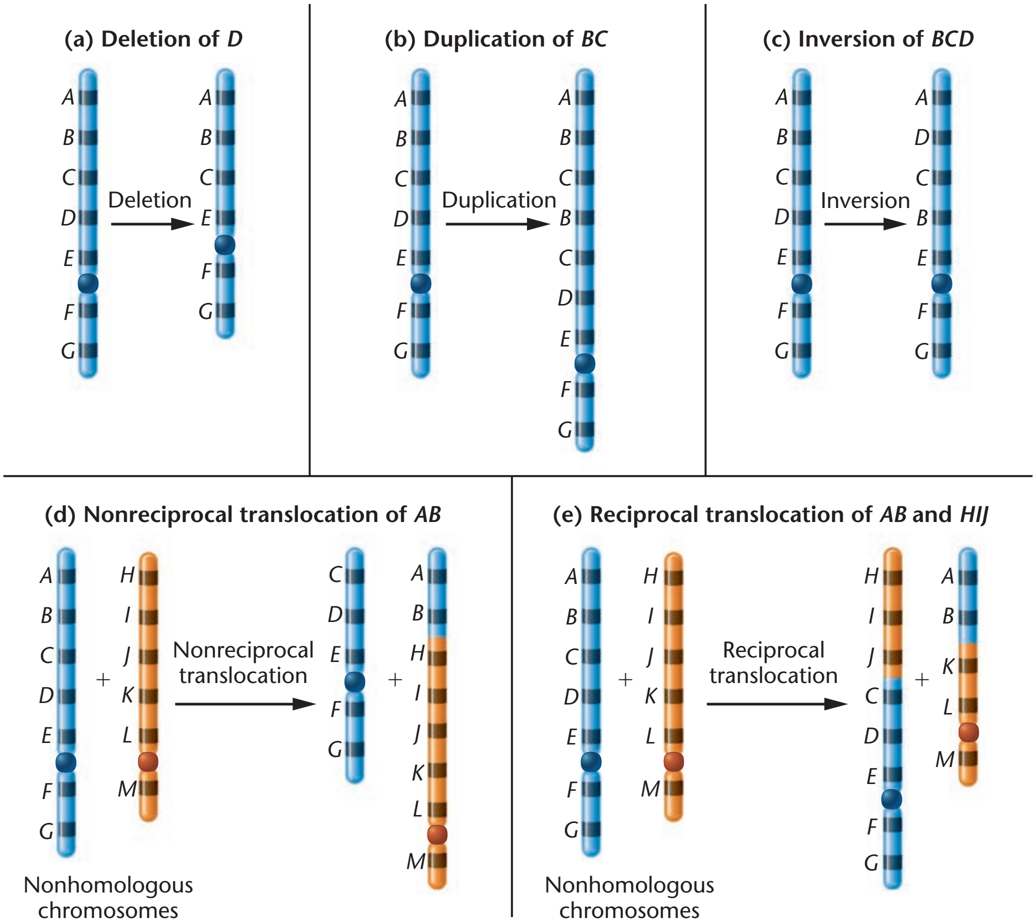


FIGURE 8.10 An overview of the five different types of gain, loss, or rearrangement of chromosome segments.

to be unaffected phenotypically. However, the unusual pairing arrangements often lead to gametes that are duplicated or deficient for some chromosomal regions. When this occurs, the offspring of "carriers" of certain aberrations have an increased probability of demonstrating phenotypic changes.

8.5 A Deletion Is a Missing Region of a Chromosome

When a chromosome breaks in one or more places and a portion of it is lost, the missing piece is called a **deletion** (or a **deficiency**). The deletion can occur either near one end or within the interior of the chromosome. These are **terminal** and **intercalary deletions**, respectively [Figure 8.11(a) and (b)]. The portion of the chromosome that retains the centromere region is usually maintained when the cell divides, whereas the segment without the centromere is eventually lost in progeny cells following mitosis or meiosis. For synapsis to occur between a chromosome with a large intercalary deletion and a normal homolog, the unpaired region of the normal homolog must "buckle out" into a **deletion**, or **compensation, loop** [Figure 8.11(c)].

If only a small part of a chromosome is deleted, the organism might survive. However, a deletion of a portion of a chromosome need not be very great before the effects become severe. We see an example of these effects in the following discussion of the cri du chat syndrome in humans. If even more genetic information is lost as a result of a deletion, the aberration is often lethal, in which case the chromosome mutation never becomes available for study.

Cri du Chat Syndrome in Humans

In humans, the **cri du chat syndrome** results from the deletion of a small terminal portion of chromosome 5. It might be considered a case of *partial monosomy,* but since the region that is missing is so small, it is better referred to as a **segmental deletion**. This syndrome was first reported by Jérôme LeJeune in 1963, when he described the clinical symptoms, including an eerie cry similar to the meowing of a cat, after which the syndrome is named. This syndrome is associated with the loss of a small, variable part of the short

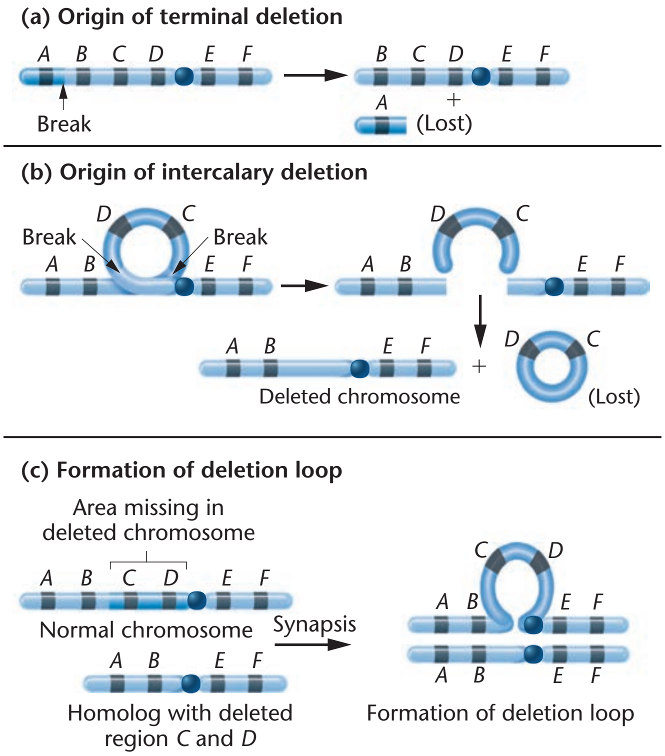

(a) Origin of terminal deletion
A B C D E F
Break
B C D E F
A
+
(Lost)
(b) Origin of intercalary deletion
D C
Break
Break
A B E F
A B
D C
E F
A B E F
Deleted chromosome
+
D C
(Lost)
(c) Formation of deletion loop
Area missing in deleted chromosome
A B C D E F
Normal chromosome
Synapsis
A B E F
Homolog with deleted region C and D
C D
A B E F
A B E F
Formation of deletion loop

FIGURE 8.11 Origins of (a) a terminal and (b) an intercalary deletion. In (c), pairing occurs between a normal chromosome and one with an intercalary deletion by looping out the undeleted portion to form a deletion (or compensation) loop.

arm of chromosome 5 (Figure 8.12). Thus, the genetic constitution may be designated as 46,5p−, meaning that the individual has all 46 chromosomes but that some or all of the p arm (the petite, or short, arm) of one member of the chromosome 5 pair is missing.

Infants with this syndrome may exhibit anatomic malformations, including gastrointestinal and cardiac complications, and they are often mentally retarded. Abnormal development of the glottis and larynx (leading to the characteristic cry) is typical of this syndrome.

Since 1963, hundreds of cases of cri du chat syndrome have been reported worldwide. An incidence of 1 in 25,000–50,000 live births has been estimated. Most often, the condition is not inherited but instead results from the sporadic loss of chromosomal material in gametes. The length of the short arm that is deleted varies somewhat; longer deletions appear to have a greater impact on the physical, psychomotor, and mental skill levels of those children who survive. Although the effects of the syndrome are severe, most individuals achieve motor and language skills and may be cared for at home. In 2004, it was reported that the portion of the chromosome that is missing contains the *TERT* gene, which encodes *telomerase reverse transcriptase*, an enzyme essential for the maintenance of telomeres during DNA replication (see Chapter 11). Whether the absence of this gene on one homolog is related to the multiple phenotypes of infants with cri du chat syndrome is still unknown.

8.6 A Duplication Is a Repeated Segment of a Chromosome

When any part of the genetic material—a single locus or a large piece of a chromosome—is present more than once in the genome, it is called a **duplication**. As in deletions, pairing in heterozygotes can produce a compensation loop. Duplications may arise as the result of unequal crossing over between synapsed chromosomes during meiosis (Figure 8.13) or through a replication error prior to meiosis. In the former case, both a duplication and a deletion are produced.

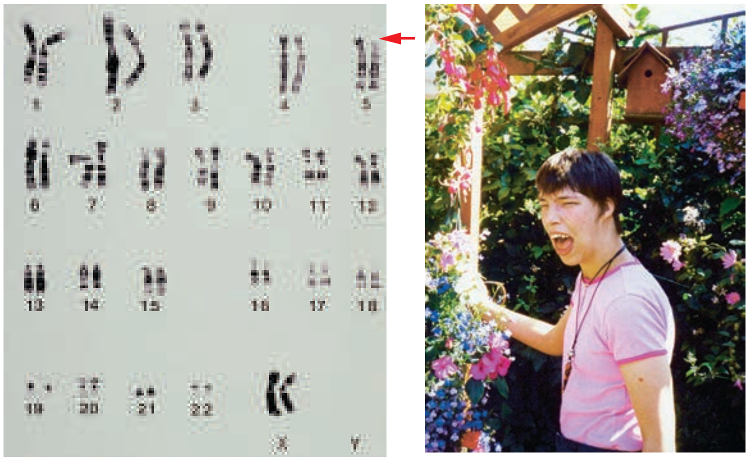


FIGURE 8.12 A representative karyotype and a photograph of a child exhibiting cri du chat syndrome (46,5p−). In the karyotype, the arrow identifies the absence of a small piece of the short arm of one member of the chromosome 5 homologs.

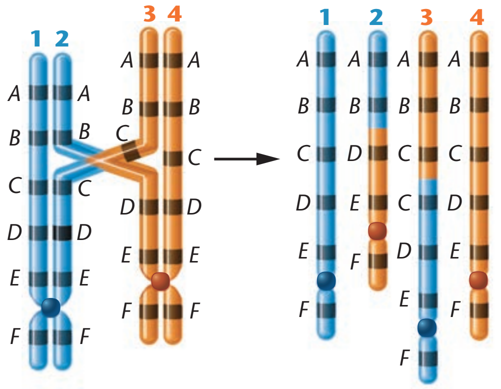


FIGURE 8.13 The origin of duplicated and deficient regions of chromosomes as a result of unequal crossing over. The tetrad on the left is mispaired during synapsis. A single crossover between chromatids 2 and 3 results in the deficient (chromosome 2) and duplicated (chromosome 3) chromosomal regions shown on the right. The two chromosomes uninvolved in the crossover event remain normal in gene sequence and content.

We consider three interesting aspects of duplications. First, they may result in gene redundancy. Second, as with deletions, duplications may produce phenotypic variation. Third, according to one convincing theory, duplications have also been an important source of genetic variability during evolution.

Gene Redundancy and Amplification—Ribosomal RNA Genes

Duplication of chromosomal segments has the potential to amplify the number of copies of individual genes. This has clearly been the case with the gene-encoding ribosomal RNA, which is needed in abundance in the ribosomes of all cells to support protein synthesis. We might hypothesize that a single copy of the gene-encoding rRNA is inadequate in many cells that demonstrate intense metabolic activity. Studies using the technique of molecular hybridization, which enables us to determine the percentage of the genome that codes for specific RNA sequences, show that our hypothesis is correct. Indeed, multiple copies of genes code for rRNA. Such DNA is called **rDNA**, and the general phenomenon is referred to as **gene redundancy**. For example, in the common intestinal bacterium *Escherichia coli* (*E. coli*), about 0.7 percent of the haploid genome consists of rDNA—the equivalent of seven copies of the gene. In *Drosophila melanogaster,* 0.3 percent of the haploid genome, equivalent to 130 gene copies, consists of rDNA. Although the presence of multiple copies of the same gene is not restricted to those coding for rRNA, we will focus on those genes in this section.

Interestingly, in some cells, particularly oocytes, even the normal redundancy of rDNA is insufficient to provide adequate amounts of rRNA needed to construct ribosomes. Oocytes store abundant nutrients and huge quantities of ribosomes, for use by the embryo during early development. More ribosomes are included in oocytes than in any other cell type. By considering how the amphibian *Xenopus laevis* acquires this abundance of ribosomes, we shall see a second way in which the amount of rRNA is increased. This phenomenon is called **gene amplification**.

The genes that code for rRNA are located in an area of the chromosome known as the **nucleolar organizer region (NOR)**. The NOR is intimately associated with the nucleolus, which is a processing center for ribosome production. Molecular hybridization analysis has shown that each NOR in the frog *Xenopus* contains the equivalent of 400 redundant gene copies coding for rRNA. Even this number of genes is apparently inadequate to synthesize the vast amount of ribosomes that must accumulate in the amphibian oocyte to support development following fertilization.

To further amplify the number of rRNA genes, the rDNA is selectively replicated, and each new set of genes is released from its template. Because each new copy is equivalent to one NOR, multiple small nucleoli are formed in the oocyte. As many as 1500 of these “micronucleoli” have been observed in a single *Xenopus* oocyte. If we multiply the number of micronucleoli (1500) by the number of gene copies in each NOR (400), we see that amplification in *Xenopus* oocytes can result in over half a million gene copies! If each copy is transcribed only 20 times during the maturation of the oocyte, in theory, sufficient copies of rRNA are produced to result in well over 12 million ribosomes.

The *Bar* Mutation in *Drosophila*

Duplications can cause phenotypic variation that might at first appear to be caused by a simple gene mutation. The *Bar*-eye phenotype in *Drosophila* (**Figure 8.14**) is a classic example. Instead of the normal oval-eye shape, *Bar*-eyed flies have narrow, slitlike eyes. This phenotype is inherited in the same way as a dominant X-linked mutation.

In the early 1920s, Alfred H. Sturtevant and Thomas H. Morgan discovered and investigated this “mutation.” Normal wild-type females (B^+/B^+) have about 800 facets in each eye. Heterozygous females (B/B^+) have about 350 facets, while homozygous females (*B*/*B*) average only about 70 facets. Females were occasionally recovered with even fewer facets and were designated as *double Bar* (B^D/B^+).

About 10 years later, Calvin Bridges and Herman J. Muller compared the polytene X chromosome banding pattern of the *Bar* fly with that of the wild-type fly. These chromosomes contain specific banding patterns that have been well categorized into regions. Their studies revealed that one copy of the region designated as 16A is present on both X chromosomes of wild-type flies but that this region

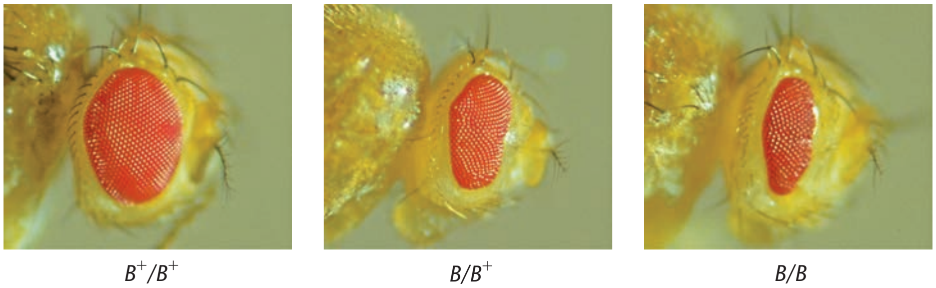


FIGURE 8.14 *Bar*-eye phenotypes in contrast to the wild-type eye in *Drosophila*.

was duplicated in *Bar* flies and triplicated in *double Bar* flies. These observations provided evidence that the *Bar* phenotype is not the result of a simple chemical change in the gene but is instead a duplication.

The Role of Gene Duplication in Evolution

During the study of evolution, it is intriguing to speculate on the possible mechanisms of genetic variation. The origin of unique gene products present in more recently evolved organisms but absent in ancestral forms is a topic of particular interest. In other words, how do "new" genes arise? As we will see below, the process of gene duplication is hypothesized to be the major source of new genes, as proposed in 1970 by Susumu Ohno in his provocative monograph, *Evolution by Gene Duplication.* Ohno's thesis is based on the supposition that the products of many genes, present as only a single copy in the genome, are indispensable to the survival of members of any species during evolution. Therefore, unique genes are not free to accumulate mutations sufficient to alter their primary function and give rise to new genes.

However, if an essential gene is duplicated in the germ line, major mutational changes in this extra copy will be tolerated in future generations because the original gene provides the genetic information for its essential function. The duplicated copy will be free to acquire many mutational changes over extended periods of time. Over short intervals, the new genetic information may be of no practical advantage. However, over long evolutionary periods, the duplicated gene may change sufficiently so that its product assumes a divergent role in the cell. The new function may impart an "adaptive" advantage to organisms, enhancing their fitness. Ohno has outlined a mechanism through which sustained genetic variability may have originated.

Ohno's thesis is supported by the discovery of genes that have a substantial amount of their DNA sequence in common, but whose gene products are distinct. The genes encoding the digestive enzymes *trypsin* and *chymotrypsin* are examples, as are those that encode the respiratory molecules *myoglobin* and the various forms of *hemoglobin.* We conclude that the genes arose from a common ancestral gene through duplication. During evolution, the related genes diverged sufficiently that their products became unique.

Other support includes the presence of **gene families**—groups of contiguous genes whose products perform the same, or very similar functions. Again, members of a family show DNA sequence homology sufficient to conclude that they share a common origin and arose through the process of gene duplication. One of the most interesting supporting examples is the case of the *SRGAP2* gene in primates. This gene is known to be involved in the development of the brain. Humans have at least four similar copies of the gene, while all nonhuman primates have only a single copy. Several duplication events can be traced back to 3.4 million years ago, to 2.4 million years ago, and finally to 1 million years ago, resulting in distinct forms of *SRGAP2* labeled A–D. These evolutionary periods coincide with the emergence of the human lineage in primates. The function of these genes has now been related to the enhancement of dendritic spines in the brain, which is believed to contribute to the evolution of expanded brain function in humans, including the development of language and social cognition.

Other examples of gene families arising from duplication during evolution include the various types of human hemoglobin polypeptide chains (Chapter 14), the immunologically important T-cell receptors and antigens encoded by the major histocompatibility complex, and the clusters of multiple *Hox* genes that are important during development in vertebrates (see Chapter 23).

Duplications at the Molecular Level: Copy Number Variations (CNVs)

As we entered the era of genomics and became capable of sequencing entire genomes (see Chapters 20 and 21), it was quickly realized that duplications of large DNA sequences, most often involving thousands of base pairs, occur on a regular basis. When individuals in the same species are compared, the number of copies of any given sequence is found to vary—sometimes there are larger numbers, and in other cases, copies have been deleted, resulting in smaller

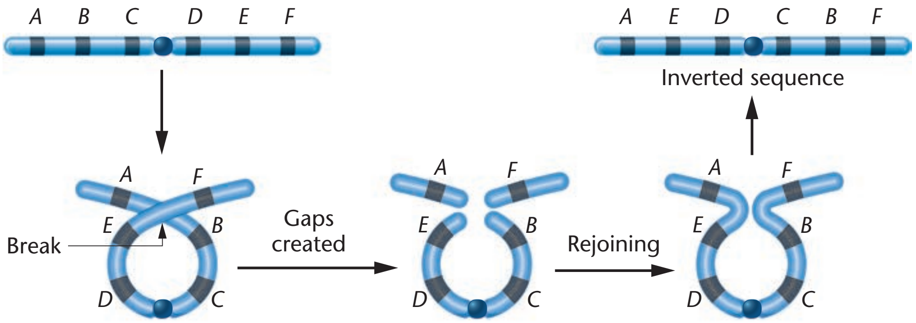


FIGURE 8.15 One possible origin of a pericentric inversion.

numbers. These variations, because they represent *quantitative differences in the number of DNA sequences,* are termed **copy number variations (CNVs)** and are found in both coding and noncoding regions of the genome.

CNVs are of major interest in genetics because they are now believed to play crucial roles in the expression of many of our individual traits, in both normal and diseased individuals. Currently, when CNVs of sizes ranging from 50 bp to 3 Mb are considered, it is estimated that they occupy between 5–10 percent of the human genome. Current studies have focused on finding associations with human diseases. CNVs appear to have both positive and negative associations with many diseases in which the genetic basis is not yet fully understood. For example, pathogenic CNVs have been associated with autism and other neurological disorders, and with cancer. Additionally, CNVs are suspected to be associated with Type I diabetes and cardiovascular disease.

In some cases, entire gene sequences are duplicated and impact individuals. For example, a higher-than-average copy number of the gene *CCL3L1* imparts an HIV-suppressive effect during viral infection, diminishing the progression to AIDS. Another finding has associated specific mutant CNV sites with certain subset populations of individuals with lung cancer—the greater number of copies of the *EGFR* (*Epidermal Growth Factor Receptor*) gene, the more responsive are patients with non-small-cell lung cancer to treatment. Finally, the greater the reduction in the copy number of the gene designated *DEFB,* the greater the risk of developing Crohn's disease, a condition affecting the colon. Relevant to this chapter, these findings reveal that duplications and deletions are no longer restricted to textbook examples of these chromosomal mutations. We will return to this interesting topic later in the text (see Chapter 21), when genomics is discussed in detail.

8.7 Inversions Rearrange the Linear Gene Sequence

The **inversion**, another class of structural variation, is a type of chromosomal aberration in which a segment of a chromosome is turned around 180 degrees within a chromosome. An inversion does not involve a loss of genetic information but simply rearranges the linear gene sequence. An inversion requires breaks at two points along the length of the chromosome and subsequent reinsertion of the inverted segment. **Figure 8.15** illustrates how an inversion might arise. By forming a chromosomal loop prior to breakage, the newly created "sticky ends" are brought close together and rejoined.

The inverted segment may be short or quite long and may or may not include the centromere. If the centromere is not part of the rearranged chromosome segment, it is a **paracentric inversion**. If the centromere is part of the inverted segment, it is described as a **pericentric inversion**, which is the type shown in Figure 8.15.

Although inversions appear to have a minimal impact on the individuals bearing them, their consequences are of great interest to geneticists. Organisms that are heterozygous for inversions may produce aberrant gametes that have a major impact on their offspring.

Consequences of Inversions during Gamete Formation

If only one member of a homologous pair of chromosomes has an inverted segment, normal linear synapsis during meiosis is not possible. Organisms with one inverted chromosome and one noninverted homolog are called **inversion heterozygotes**. Two such chromosomes in meiosis can be paired only if they form an **inversion loop** (**Figure 8.16**).

If crossing over does not occur within the inverted segment of the inversion loop, the homologs will segregate, which results in two normal and two inverted chromatids that are distributed into gametes. However, if crossing over does occur within the inversion loop, abnormal chromatids are produced. The effect of a single crossover (SCO) event within a paracentric inversion is diagrammed in **Figure 8.16(a)**.

In any meiotic tetrad, a single crossover between nonsister chromatids produces two parental chromatids and two recombinant chromatids. When the crossover occurs within a paracentric inversion, however, one recombinant **dicentric chromatid** (two centromeres) and one recombinant **acentric chromatid** (lacking a centromere) are produced. Both contain duplications and deletions of chromosome segments as well. During anaphase, an acentric

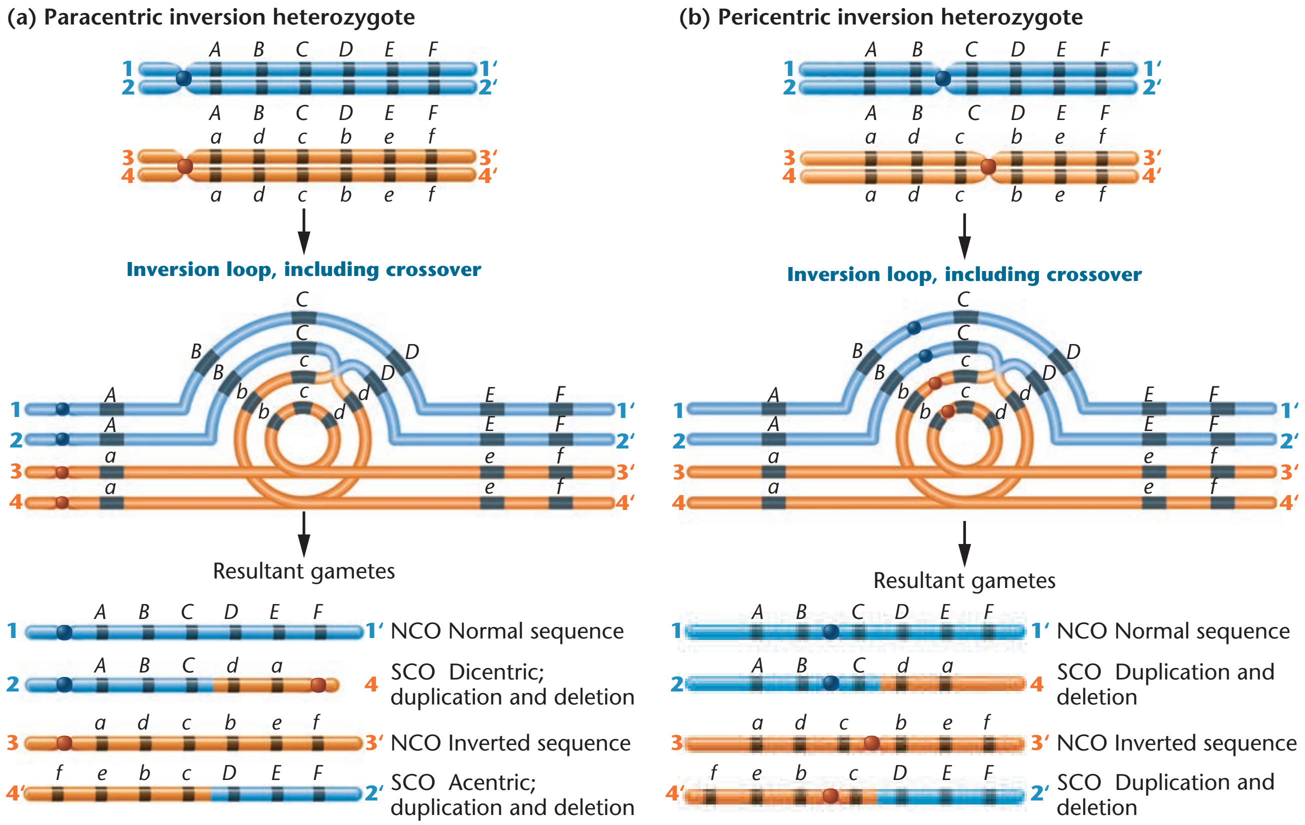

(a) Paracentric inversion heterozygote
1 A B C D E F 1′
2 A B C D E F 2′
3 a d c b e f 3′
4 a d c b e f 4′
Inversion loop, including crossover
Resultant gametes
1 A B C D E F 1′ NCO Normal sequence
2 A B C d a 4 SCO Dicentric; duplication and deletion
3 a d c b e f 3′ NCO Inverted sequence
4′ f e b c D E F 2′ SCO Acentric; duplication and deletion
(b) Pericentric inversion heterozygote
1 A B C D E F 1′
2 A B C D E F 2′
3 a d c b e f 3′
4 a d c b e f 4′
Inversion loop, including crossover
Resultant gametes
1 A B C D E F 1′ NCO Normal sequence
2 A B C d a 4 SCO Duplication and deletion
3 a d c b e f 3′ NCO Inverted sequence
4′ f e b c D E F 2′ SCO Duplication and deletion

FIGURE 8.16 (a) The effects of a single crossover (SCO) within an inversion loop in a paracentric inversion heterozygote, where two altered chromosomes are produced, one acentric and one dicentric. Both chromosomes also contain duplicated and deficient regions. (b) The effects of a crossover in a pericentric inversion heterozygote, where two altered chromosomes are produced, both with duplicated and deficient regions.

chromatid moves randomly to one pole or the other or may be lost, while a dicentric chromatid is pulled in two directions. This polarized movement produces *dicentric bridges* that are cytologically recognizable. A dicentric chromatid usually breaks at some point so that part of the chromatid goes into one gamete and part into another gamete during the reduction divisions. Therefore, gametes containing either recombinant chromatid are deficient in genetic material. When such a gamete participates in fertilization, the zygote most often develops abnormally, if at all.

A similar chromosomal imbalance is produced as a result of a crossover event between a chromatid bearing a pericentric inversion and its noninverted homolog, as shown in **Figure 8.16(b)**. The recombinant chromatids that are directly involved in the exchange have duplications and deletions. In plants, gametes receiving such aberrant chromatids fail to develop normally, leading to aborted pollen or ovules. Thus, lethality occurs prior to fertilization, and inviable seeds result. In animals, the gametes have developed prior to the meiotic error, so fertilization is more likely to occur in spite of the chromosome error. However, the end result is the production of inviable embryos following fertilization. In both cases, viability is reduced.

Because offspring bearing crossover gametes are inviable and not recovered, it *appears* as if the inversion suppresses crossing over. Actually, in inversion heterozygotes, the inversion has the effect of *suppressing the recovery of crossover products* when chromosome exchange occurs within the inverted region. If crossing over always occurred within a paracentric or pericentric inversion, 50 percent of the gametes would be ineffective. The viability of the resulting zygotes is therefore greatly diminished. Furthermore, up to one-half of the viable gametes have the inverted chromosome, and the inversion will be perpetuated within the species. The cycle will be repeated continuously during meiosis in future generations.

Evolutionary Advantages of Inversions

Because recovery of crossover products is suppressed in inversion heterozygotes, groups of specific alleles at adjacent loci within inversions may be preserved from generation to generation. If the alleles of the involved genes confer a survival advantage on organisms maintaining them, the inversion is beneficial to the evolutionary survival of the species. For example, if a set of alleles *ABcDef* is more adaptive than sets *AbCdeF* or *abcdEF*, effective gametes will contain this favorable set of genes, undisrupted by crossing over.

In laboratory studies, the same principle is applied using **balancer chromosomes**, which contain inversions. When an organism is heterozygous for a balancer chromosome, desired sequences of alleles are preserved during experimental work.

NOW SOLVE THIS

8.3 What is the effect of a rare double crossover (a) within a chromosome segment that is heterozygous for a pericentric inversion; and (b) within a segment that is heterozygous for a paracentric inversion?

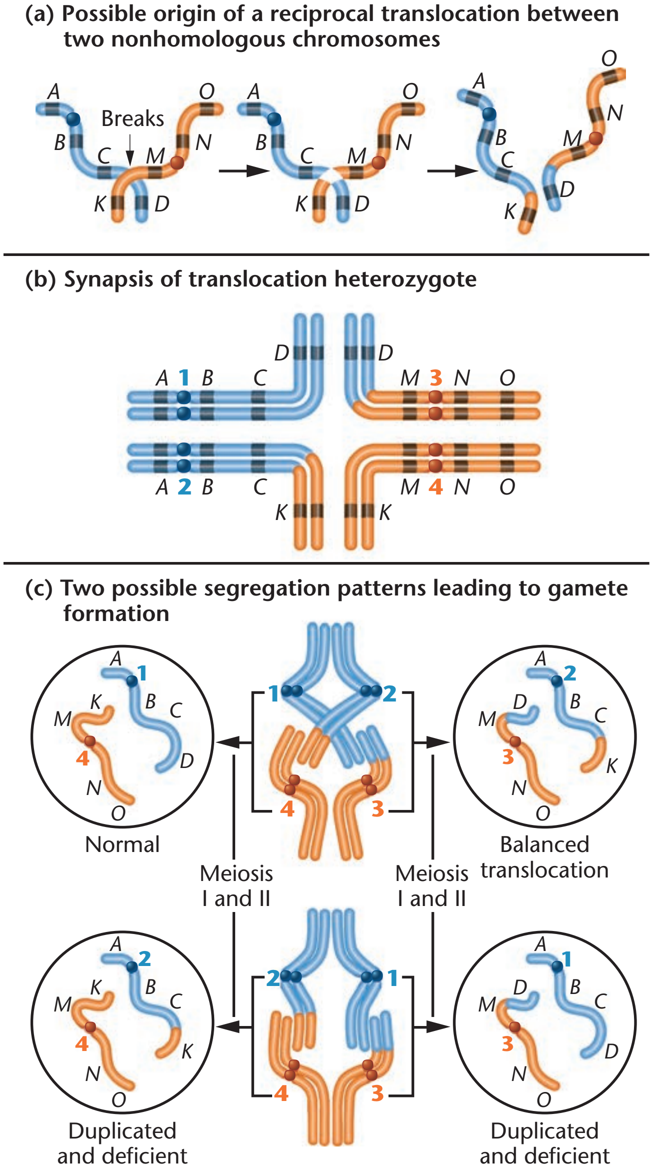
- **HINT:** *This problem involves an understanding of how homologs synapse in the presence of a heterozygous inversion, as well as the distinction between pericentric and paracentric inversions. The key to its solution is to draw out the tetrad and follow the chromatids undergoing a double crossover.*

8.8 Translocations Alter the Location of Chromosomal Segments in the Genome

Translocation, as the name implies, is the movement of a chromosomal segment to a new location in the genome. Reciprocal translocation, for example, involves the exchange of segments between two nonhomologous chromosomes. The least complex way for this event to occur is for two nonhomologous chromosome arms to come close to each other so that an exchange is facilitated. **Figure 8.17(a)** shows a simple reciprocal translocation in which only two breaks are required. If the exchange includes internal chromosome segments, four breaks are required, two on each chromosome.

The genetic consequences of reciprocal translocations are, in several instances, similar to those of inversions. For example, genetic information is not lost or gained. Rather, there is only a rearrangement of genetic material. The presence of a translocation does not, therefore, directly alter the viability of individuals bearing it.

Homologs that are heterozygous for a reciprocal translocation undergo unorthodox synapsis during meiosis. As shown in **Figure 8.17(b)**, pairing results in a cruciform, or crosslike, configuration. As with inversions, some of the gametes produced are genetically unbalanced as a result of an unusual alignment during meiosis. In the case of translocations, however, aberrant gametes are not necessarily the result of crossing over. To see how unbalanced gametes are produced, focus on the homologous centromeres in **Figure 8.17(b)** and **(c)**. According to the principle of independent assortment, the chromosome containing centromere 1 migrates randomly toward one pole of the spindle during the first meiotic anaphase; it travels along with *either* the chromosome having centromere 3 *or* the chromosome having centromere 4. The chromosome with centromere 2 moves to the other pole, along with the chromosome containing *either* centromere 3 *or* centromere 4. This results in four potential meiotic products. The 1,4 combination contains chromosomes that are not involved in the translocation and are normal. The 2,3 combination, however, contains the translocated chromosomes, but


(a) Possible origin of a reciprocal translocation between two nonhomologous chromosomes

(b) Synapsis of translocation heterozygote

(c) Two possible segregation patterns leading to gamete formation

FIGURE 8.17 (a) Possible origin of a reciprocal translocation. (b) Synaptic configuration formed during meiosis in an individual that is heterozygous for the translocation. (c) Two possible segregation patterns, one of which leads to a normal and a balanced gamete (called alternate segregation) and one that leads to gametes containing duplications and deficiencies (called adjacent segregation-1).

these contain a complete complement of genetic information and are balanced. When this result is achieved [the top configuration in Figure 8.17(c)], the segregation pattern at the first meiotic division is referred to as **alternate segregation**. A second pattern [the bottom configuration in Figure 8.17(c)] produces the other two potential products, the 1,3 and 2,4 combinations, which contain chromosomes displaying duplicated and deleted (deficient) segments. This pattern is called **adjacent segregation-1.** Note that a third type of arrangement, where homologous centromeres segregate to the same pole during meiosis (called *adjacent segregation-2*), has not been included in this figure. This type of segregation has an outcome similar to adjacent segregation-1, with meiotic products containing genetically unbalanced duplicated and deleted chromosomal material.

When genetically unbalanced gametes participate in fertilization in animals, the resultant offspring do not usually survive. Fewer than 50 percent of the progeny of parents heterozygous for a reciprocal translocation survive. This condition in a parent is called **semisterility,** and its impact on the reproductive fitness of organisms plays a role in evolution. In humans, such an unbalanced condition results in partial monosomy or trisomy, leading to a variety of birth defects.

Translocations in Humans: Familial Down Syndrome

Research conducted since 1959 has revealed numerous translocations in members of the human population. One common type of translocation involves breaks at the extreme ends of the short arms of two nonhomologous acrocentric chromosomes. These small segments are lost, and the larger segments fuse at their centromeric region. This type of translocation produces a new, large submetacentric or metacentric chromosome, often called a **Robertsonian translocation**.

One such translocation accounts for cases in which Down syndrome is familial (inherited). Earlier in this chapter, we pointed out that most instances of Down syndrome are due to trisomy 21. This chromosome composition results from nondisjunction during meiosis in one parent. Trisomy accounts for over 95 percent of all cases of Down syndrome. In such instances, the chance of the same parents producing a second affected child is extremely low. However, in the remaining families with a Down child, the syndrome occurs with much greater frequency over several generations—it "runs in families."

Cytogenetic studies of the parents and their offspring from these unusual cases explain the cause of **familial Down syndrome**. Analysis reveals that one of the parents contains a **14/21 translocation** (Figure 8.18). That is, one parent has the majority of chromosome 21 translocated to one end of chromosome 14. This individual is phenotypically normal, even though he or she has only 45 chromosomes and is referred to as a *balanced translocation carrier*. During meiosis in such an individual, one of the gametes contains two copies of chromosome 21—a normal chromosome and a second copy translocated to chromosome 14. When such a gamete is fertilized by a standard haploid gamete, the resulting zygote has 46 chromosomes but three copies of chromosome 21. These individuals exhibit Down syndrome. Other potential surviving offspring contain either the standard diploid genome (without a translocation) or the balanced translocation like the parent. Both cases result in normal individuals. In the fourth case, a monosomic individual is produced, which is lethal. Although not illustrated, adjacent-2 segregation is also thought to occur, but rarely. Such gametes are unbalanced, and upon fertilization, lethality occurs.

The above findings have allowed geneticists to resolve the seeming paradox of an inherited trisomic phenotype in an individual with an apparent diploid number

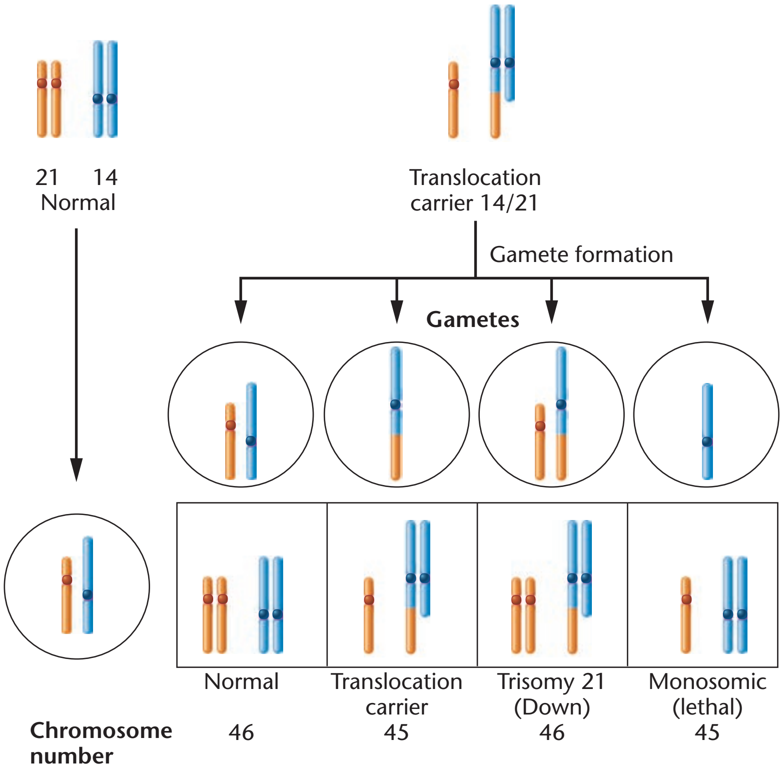


FIGURE 8.18 Chromosomal involvement and translocation in familial Down syndrome.

of chromosomes. Interestingly, the "carrier," who has 45 chromosomes and exhibits a normal phenotype, does not contain the complete diploid amount of genetic material. A small region is lost from both chromosomes 14 and 21 during the translocation event. This occurs because the ends of both chromosomes have broken off prior to their fusion. These specific regions are known to be two of many chromosomal locations housing multiple copies of the genes encoding rRNA, the major component of ribosomes. Despite the loss of up to 20 percent of these genes, the carrier is unaffected.

8.9 Fragile Sites in Human Chromosomes Are Susceptible to Breakage

We conclude this chapter with a brief discussion of the results of an intriguing discovery made around 1970 during observations of metaphase chromosomes prepared following human cell culture. In cells derived from certain individuals, a specific area along one of the chromosomes failed to stain, giving the appearance of a gap. In other individuals whose chromosomes displayed such morphology, the gaps appeared at other positions within the set of chromosomes. Such areas eventually became known as **fragile sites**, since they appeared to be susceptible to chromosome breakage when cultured in the absence of certain chemicals such as folic acid, which is normally present in the culture medium.

The cause of the fragility at these sites is unknown. However, since they represent points along the chromosome that are susceptible to breakage, these sites may indicate regions where the chromatin is not tightly coiled. Note that even though almost all studies of fragile sites have been carried out *in vitro* using mitotically dividing cells, interest in them increased when clear associations were established between several of these sites and a corresponding altered phenotype, including mental retardation and cancer.

Fragile-X Syndrome

While most fragile sites do not appear to be associated with any clinical syndrome, individuals bearing a folate-sensitive site on the X chromosome (Figure 8.19) exhibit the **fragile-X syndrome** (**FXS**), the most common form of inherited mental retardation. This syndrome affects about 1 in 4000 males and 1 in 8000 females. Since affected females usually carry only one fragile X chromosome, the disorder is considered a dominant trait. Fortunately, penetrance is not complete, and the trait is fully expressed in only about 30 percent of fragile-X–bearing females and 80 percent of fragile-X–bearing males. In addition to mental retardation, affected males have characteristic long, narrow faces with protruding chins, enlarged ears, and increased testicular size.

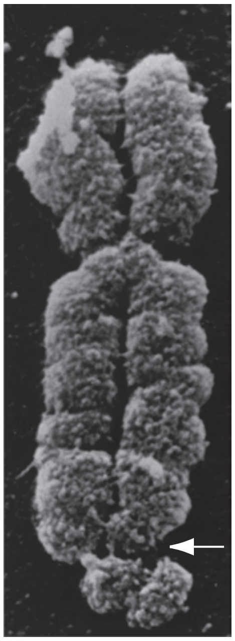


FIGURE 8.19 A human fragile X chromosome. The "gap" region (near the bottom of the chromosome) is associated with the fragile-X syndrome.

A gene that spans the fragile site, *FMR1*, is now known to be responsible for this syndrome. It is one of many genes in which a sequence of three nucleotides is repeated many times, expanding the size of the gene. Such **trinucleotide repeats** are also characteristic of other human disorders, including Huntington disease and myotonic dystrophy. In *FMR1,* the trinucleotide sequence CGG is repeated in an untranslated area adjacent to the coding sequence of the gene (called the "upstream" region). The number of repeats varies within the human population, and a high number correlates directly with expression of FXS. Normal individuals have between 6 and 54 repeats, whereas those with 55 to 230 repeats are unaffected but are considered "carriers" of the disorder. More than 230 repeats lead to expression of the syndrome.

The primary impact of the increased number of repeats in FXS is the loss of expression of *FMR1*. It is thought that, once the gene contains this increased number of repeats, it becomes chemically modified so that the bases within and around the repeats are methylated, which inactivates the gene. The normal product of the gene is an RNA-binding protein, FMRP. Evidence is now accumulating that directly links the absence of FMRP in the brain with the cognitive defects associated with the syndrome. RNA-binding proteins and their role in gene regulation are discussed in more detail in Chapter 18.

From a genetic standpoint, an interesting aspect of FXS is the instability of the CGG repeats. An individual with 6 to 54 repeats transmits a gene containing the same number of copies to his or her offspring. However, carrier individuals with 55 to 230 repeats, though not at risk to develop the syndrome, may transmit to their offspring a gene with an increased number of repeats. The number of repeats increases in future generations, demonstrating the phenomenon known as **genetic anticipation**. Once the threshold of 230 repeats is exceeded, retardation becomes more severe in each successive generation as the number of trinucleotide repeats increases. Interestingly, expansion from the carrier status (55 to 230 repeats) to the syndrome status (over 230 repeats) occurs only during the transmission of the gene by the maternal parent, not by the paternal parent. Thus, a "carrier" male may transmit a stable chromosome to his daughter, who may subsequently transmit an unstable chromosome with an increased number of repeats to her offspring. Their grandfather was the source of the original chromosome.

The Link between Fragile Sites and Cancer

While the study of the FXS first brought unstable chromosome regions to the attention of geneticists, a link between an autosomal fragile site and lung cancer was reported in 1996 by Carlo Croce, Kay Huebner, and their colleagues. They demonstrated that a gene, *FHIT* (standing for *f*ragile *hi*stidine *t*riad), located within the well-defined fragile site designated as *FRA3B* on the p arm of chromosome 3, is often altered or missing in cells taken from tumors of individuals with lung cancer. More extensive studies have revealed that the normal protein product of this gene is absent in cells of many other cancers, including those of the esophagus, breast, cervix, liver, kidney, pancreas, colon, and stomach. Genes such as *FHIT* that are located within fragile regions undoubtedly have an increased susceptibility to mutations and deletions.

The study of this and still other fragile sites is but one example of how chromosomal abnormalities of many sorts are linked to cancer. We will return to this discussion in Chapter 24.

GENETICS, ETHICS, AND SOCIETY

Down Syndrome and Prenatal Testing—The New Eugenics?

Down syndrome is the most common chromosomal abnormality seen in newborn babies. Prenatal diagnostic tests for Down syndrome have been available for decades, especially to older pregnant women who have an increased risk of bearing a child with Down syndrome. Scientists estimate that there is an abortion rate of about 30 percent for fetuses that test positive for Down syndrome in the United States, and rates of up to 85 percent in other parts of the world, such as Taiwan and France.

Many people agree that it is morally acceptable to prevent the birth of a genetically abnormal fetus. However, many others argue that prenatal genetic testing, with the goal of eliminating congenital disorders, is unethical. In addition, some argue that prenatal genetic testing followed by selective abortion is eugenic. How does eugenics apply, if at all, to screening for Down syndrome and other human genetic defects?

The term *eugenics* was first defined by Francis Galton in 1883 as "the science which deals with all influences that improve the inborn qualities of a race; also with those that develop them to the utmost advantage." Galton believed that human traits such as intelligence and personality were hereditary and that humans could selectively mate with each other to create gifted groups of people—analogous to the creation of purebred dogs with specific traits. Galton did not propose coercion but thought that people would voluntarily select mates in order to enhance particular genetic outcomes for their offspring.

In the early to mid-twentieth century, countries throughout the world adopted eugenic policies with the aim of enhancing desirable human traits (positive eugenics) and eliminating undesirable ones (negative eugenics). Many countries, including Britain, Canada, and the United States, enacted compulsory sterilization programs for the "feebleminded," mentally ill, and criminals. The eugenic policies of Nazi Germany were particularly infamous, resulting in forced human genetic experimentation and the slaughter of tens of thousands of disabled people. The eugenics movement was discredited after World War II, and the evils perpetuated in its name have tainted the term *eugenics* ever since.

Given the history of the eugenics movement, is it fair to use the term

(continued)

Genetics, Ethics, and Society, continued

eugenics when we speak about genetic testing for Down syndrome and other genetic disorders? Some people argue that it is not eugenic to select for healthy children because there is no coercion, the state is not involved, and the goal is the elimination of suffering. Others point out that such voluntary actions still constitute eugenics, since they involve a form of bioengineering for "better" human beings.

Now that we are entering an era of unprecedented knowledge about our genomes and our predisposition to genetic disorders, we must make decisions about whether our attempts to control or improve human genomes are ethical and what limits we should place on these efforts. The story of the eugenics movement provides us with a powerful cautionary tale about the potential misuses of genetic information.

Your Turn

Take time, individually or in groups, to consider the following questions. Investigate the references and links to help you discuss some of the ethical issues surrounding genetic testing and eugenics.

1. Do you think that modern prenatal and preimplantation genetic testing followed by selective abortion is eugenic? Why or why not?

For background on these questions, see McCabe, L., and McCabe, E. (2011). Down syndrome: Coercion and eugenics. *Genet. Med.* 13:708–710. *Another useful discussion can be found in* Wilkinson, S., (2015). Prenatal screening, reproductive choice, and public health. *Bioethics* 29:26–35.

2. If genetic technologies were more advanced than today, and you could choose the traits of your children, would you take advantage of that option? Which traits would you choose—height, weight, intellectual abilities, athleticism, artistic talents? If so, would this be eugenic? Would it be ethical?

To read about similar questions answered by groups of Swiss law and medical students, read Elger, B., and Harding, T., (2003). Huntington's disease: Do future physicians and lawyers think eugenically? *Clin. Genet.* 64:327–338.

CASE STUDY Fish tales

Controlling the overgrowth of invasive aquatic vegetation is a significant problem in the waterways of most U.S. states. Originally, herbicides and dredging were used for control, but in 1963, diploid Asian carp were introduced in Alabama and Arkansas. Unfortunately, through escapes and illegal introductions, the carp spread rapidly and became serious threats to aquatic ecosystems in 45 states. Beginning in 1983, many states began using triploid, sterile grass carp as an alternative, because of their inability to reproduce, their longevity, and their voracious appetite. On the other hand, this genetically modified exotic species, if not used properly, can reduce or eliminate desirable plants and outcompete native fish, causing more damage than good. The use of one exotic species to control other exotic species has had a problematic history across the globe, generating controversy and criticism. Newer methods for genetic modification of organisms to achieve specific outcomes will certainly become more common in the future and raise several interesting questions.

1. Why would the creation and use of a tetraploid carp species be unacceptable in the above situation?
2. If you were a state official in charge of a particular waterway, what questions would you ask before approving the use of a laboratory-produced, triploid species in this waterway?
3. What ethical responsibilities accompany the ecological and economic risks and benefits of releasing exotic species into the environment? Who pays the costs if ecosystems and food supplies are damaged?

See Seastedt, T. R. (2015). Biological control of invasive plant species: A reassessment for the Anthropocene. *New Phytologist* 205:490–502.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Alterations of the precise diploid content of chromosomes are referred to as chromosomal aberrations or chromosomal mutations.
2. Studies of monosomic and trisomic disorders are increasing our understanding of the delicate genetic balance that is essential for normal development.
3. When more than two haploid sets of chromosomes are present, these may be derived from the same or different species, the basis of autopolyploidy and allotetraploidy, respectively.
4. Deletions or duplications of segments of a gene or a chromosome may be the source of mutant phenotypes, such as cri du chat syndrome in humans and Bar eyes in *Drosophila,* while duplications can be particularly important as a source of amplified or new genes.
5. Inversions and translocations may initially cause little or no loss of genetic information or deleterious effects. However, heterozygous combinations of the involved chromosome segments may result in genetically abnormal gametes following meiosis, with lethality or inviability often ensuing.
6. Fragile sites in human mitotic chromosomes have sparked research interest because one such site on the X chromosome is associated with the most common form of inherited mental retardation, while other autosomal sites have been linked to various forms of cancer.

INSIGHTS AND SOLUTIONS

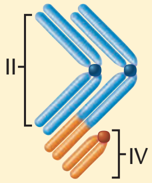
1. In a cross using maize that involves three genes, *a, b,* and *c,* a heterozygote (*abc*/+++) is testcrossed to *abc*/*abc*. Even though the three genes are separated along the chromosome, thus predicting that crossover gametes and the resultant phenotypes should be observed, only two phenotypes are recovered: *abc* and +++. In addition, the cross produced significantly fewer viable plants than expected. Can you propose why no other phenotypes were recovered and why the viability was reduced?

   Solution:
   One of the two chromosomes may contain an inversion that overlaps all three genes, effectively precluding the recovery of any "crossover" offspring. If this is a paracentric inversion and the genes are clearly separated (ensuring that a significant number of crossovers occurs between them), then numerous acentric and dicentric chromosomes will form, resulting in the observed reduction in viability.

2. A male *Drosophila* from a wild-type stock is discovered to have only seven chromosomes, whereas normally $2n = 8$. Close examination reveals that one member of chromosome IV (the smallest chromosome) is attached to (translocated to) the distal end of chromosome II and is missing its centromere, thus accounting for the reduction in chromosome number.

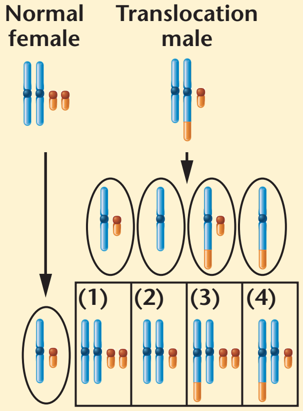
   (a) Diagram all members of chromosomes II and IV during synapsis in meiosis I.

   Solution:



   (b) If this male mates with a female with a normal chromosome composition who is homozygous for the recessive chromosome IV mutation *eyeless* (*ey*), what chromosome compositions will occur in the offspring regarding chromosomes II and IV?

   Solution:



   (c) Referring to the diagram in the solution to part (b), what phenotypic ratio will result regarding the presence of eyes, assuming all abnormal chromosome compositions survive?

   Solution:

   1. normal (heterozygous)
   2. eyeless (monosomic, contains chromosome IV from mother)
   3. normal (heterozygous; trisomic and may die)
   4. normal (heterozygous; balanced translocation)

   The final ratio is 3/4 normal: 1/4 eyeless.

Problems and Discussion Questions

1. HOW DO WE KNOW? In this chapter, we have focused on chromosomal mutations resulting from a change in number or arrangement of chromosomes. In our discussions, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) How do we know that the extra chromosome causing Down syndrome is usually maternal in origin?
(b) How do we know that human aneuploidy for each of the 22 autosomes occurs at conception, even though most often human aneuploids do not survive embryonic or fetal development and thus are never observed at birth?
(c) How do we know that specific mutant phenotypes are due to changes in chromosome number or structure?
(d) How do we know that the mutant Bar-eye phenotype in *Drosophila* is due to a duplicated gene region rather than to a change in the nucleotide sequence of a gene?

2. CONCEPT QUESTION Review the Chapter Concepts list on page 171. These all center around chromosomal aberrations that create variations from the "normal" diploid genome. Write a short essay that discusses five altered phenotypes that result from specific chromosomal aberrations.

3. Define these pairs of terms, and distinguish between them.

aneuploidy/euploidy
monosomy/trisomy
Patau syndrome/Edwards syndrome
autopolyploidy/allopolyploidy
autotetraploid/amphidiploid
paracentric inversion/pericentric inversion

4. For a species with a diploid number of 18, indicate how many chromosomes will be present in the somatic nuclei of individuals that are haploid, tetraploid, trisomic, and monosomic.

5. What evidence suggests that Down syndrome is more often the result of nondisjunction during oogenesis rather than during spermatogenesis?

6. What evidence indicates that humans with aneuploid karyotypes occur at conception but are usually inviable?

7. Contrast the fertility of an allotetraploid with an autotriploid and an autotetraploid.

8. Describe the origin of cultivated American cotton.

9. Predict how the synaptic configurations of homologous pairs of chromosomes might appear when one member is normal and the other member has sustained a deletion or duplication.

10. Inversions are said to "suppress crossing over." Is this terminology technically correct? If not, restate the description accurately.

11. Contrast the genetic composition of gametes derived from tetrads of inversion heterozygotes where crossing over occurs within a paracentric versus a pericentric inversion.

12. Human adult hemoglobin is a tetramer containing two alpha (α) and two beta (β) polypeptide chains. The α gene cluster on chromosome 16 and the β gene cluster on chromosome 11 share amino acid similarities such that 61 of the amino acids of the α-globin polypeptide (141 amino acids long) are shared in identical sequence with the β-globin polypeptide (146 amino acids long). How might one explain the existence of two polypeptides with partially shared function and structure on two different chromosomes?

13. Discuss Ohno's hypothesis on the role of gene duplication in the process of evolution. What evidence supports this hypothesis?

14. What roles have inversions and translocations played in the evolutionary process?

15. The primrose, *Primula kewensis,* has 36 chromosomes that are similar in appearance to the chromosomes in two related species, *P. floribunda* ($2n = 18$) and *P. verticillata* ($2n = 18$). How could *P. kewensis* arise from these species? How would you describe *P. kewensis* in genetic terms?

16. Certain varieties of chrysanthemums contain 18, 36, 54, 72, and 90 chromosomes; all are multiples of a basic set of nine chromosomes. How would you describe these varieties genetically? What feature do the karyotypes of each variety share? A variety with 27 chromosomes has been discovered, but it is sterile. Why?

17. *Drosophila* may be monosomic for chromosome 4, yet remain fertile. Contrast the F_1 and F_2 results of the following crosses involving the recessive chromosome 4 trait, bent bristles:

(a) monosomic IV, bent bristles × diploid, normal bristles;
(b) monosomic IV, normal bristles × diploid, bent bristles.

18. Mendelian ratios are modified in crosses involving autotetraploids. Assume that one plant expresses the dominant trait green seeds and is homozygous (*WWWW*). This plant is crossed to one with white seeds that is also homozygous (*wwww*). If only one dominant allele is sufficient to produce green seeds, predict the F_1 and F_2 results of such a cross. Assume that synapsis between chromosome pairs is random during meiosis.

19. Having correctly established the F_2 ratio in Problem 18, predict the F_2 ratio of a "dihybrid" cross involving two independently assorting characteristics (e.g., $P_1 = WWWWAAAA \times wwwwaaaa$).

20. The mutations called *bobbed* in *Drosophila* result from variable reductions (deletions) in the number of amplified genes coding for rRNA. Researchers trying to maintain *bobbed* stocks have often documented their tendency to revert to wild type in successive generations. Propose a mechanism based on meiotic recombination which could account for this reversion phenomenon. Why would wild-type flies become more prevalent in *Drosophila* cultures?

21. The outcome of a single crossover between nonsister chromatids in the inversion loop of an inversion heterozygote varies depending on whether the inversion is of the paracentric or pericentric type. What differences are expected?

22. A couple planning their family are aware that through the past three generations on the husband's side a substantial number of stillbirths have occurred and several malformed babies were born who died early in childhood. The wife has studied genetics and urges her husband to visit a genetic counseling clinic, where a complete karyotype-banding analysis is performed. Although the tests show that he has a normal complement of 46 chromosomes, banding analysis reveals that one member of the chromosome 1 pair (in group A) contains an inversion covering 70 percent of its length. The homolog of chromosome 1 and all other chromosomes show the normal banding sequence.

(a) How would you explain the high incidence of past stillbirths?
(b) What can you predict about the probability of abnormality/normality of their future children?
(c) Would you advise the woman that she will have to bring each pregnancy to term to determine whether the fetus is normal? If not, what else can you suggest?

Extra-Spicy Problems

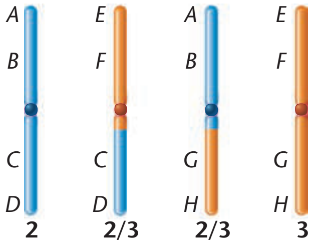
Mastering Genetics Visit for instructor-assigned tutorials and problems.

23. In a cross in *Drosophila,* a female heterozygous for the autosomally linked genes *a, b, c, d,* and *e* (*abcde*/+++++) was testcrossed with a male homozygous for all recessive alleles. Even though the distance between each of the loci was at least 3 map units, only four phenotypes were recovered, yielding the following data:

Phenotype	No. of Flies
+++++	440
a b c d e	460
++++*e*	48
a b c d +	52
	Total = 1000

Why are many expected crossover phenotypes missing? Can any of these loci be mapped from the data given here? If so, determine map distances.

24. A woman who sought genetic counseling is found to be heterozygous for a chromosomal rearrangement between the second and third chromosomes. Her chromosomes, compared to those in a normal karyotype, are diagrammed to the right.
(a) What kind of chromosomal aberration is shown?
(b) Using a drawing, demonstrate how these chromosomes would pair during meiosis. Be sure to label the different segments of the chromosomes.
(c) This woman is phenotypically normal. Does this surprise you? Why or why not? Under what circumstances might you expect a phenotypic effect of such a rearrangement?



25. The woman in Problem 24 has had two miscarriages. She has come to you, an established genetic counselor, with these questions: Is there a genetic explanation of her frequent miscarriages? Should she abandon her attempts to have a child of her own? If not, what is the chance that she could have a normal child? Provide an informed response to her concerns.

26. In a recent cytogenetic study on 1021 cases of Down syndrome, 46 were the result of translocations, the most frequent of which was symbolized as t(14;21). What does this symbol represent, and how many chromosomes would you expect to be present in t(14;21) Down syndrome individuals?

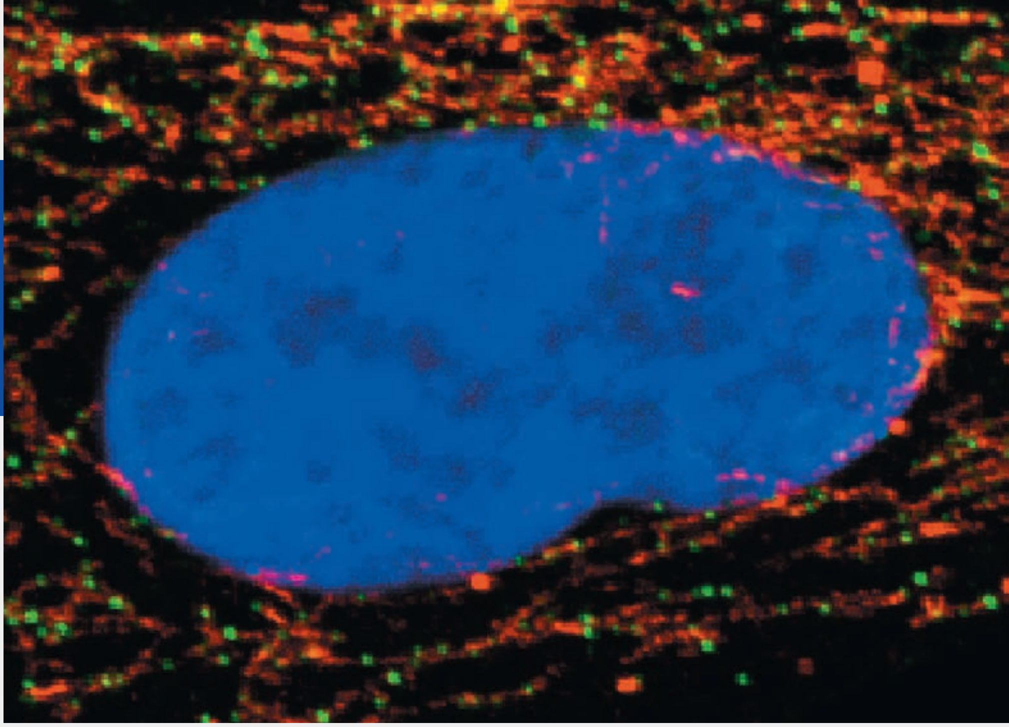
27. A boy with Klinefelter syndrome (47,XXY) is born to a mother who is phenotypically normal and a father who has the X-linked skin condition called anhidrotic ectodermal dysplasia. The mother's skin is completely normal with no signs of the skin abnormality. In contrast, her son has patches of normal skin and patches of abnormal skin.
(a) Which parent contributed the abnormal gamete?
(b) Using the appropriate genetic terminology, describe the meiotic mistake that occurred. Be sure to indicate in which division the mistake occurred.
(c) Using the appropriate genetic terminology, explain the son's skin phenotype.

28. Most cases of Turner syndrome are attributed to nondisjunction of one or more of the sex chromosomes during gametogenesis, from either the male or female parent. However, some females possess a rare form of Turner syndrome in which some of the cells of the body (somatic cells) lack an X chromosome, while other cells have the normal two X chromosomes. Often detected in blood and/or skin cells, such individuals with mosaic Turner syndrome may exhibit relatively mild symptoms. An individual may be specified as 45,X(20)/46,XX(80) if, for example, 20 percent of the cells examined were X monosomic. How might mitotic events cause such mosaicism, and what parameter(s) would likely determine the percentages and distributions of X0 cells?

29. A 3-year-old child exhibited some early indication of Turner syndrome, which results from a 45,X chromosome composition. Karyotypic analysis demonstrated two cell types: 46,XX (normal) and 45,X. Propose a mechanism for the origin of this mosaicism.

30. A normal female is discovered with 45 chromosomes, one of which exhibits a Robertsonian translocation containing most of chromosomes 15 and 21. Discuss the possible outcomes in her offspring when her husband contains a normal karyotype.

9



Extranuclear Inheritance

A mammalian cell demonstrating the presence of mtDNA (stained green), which is located in the nucleoids in the tubular mitochondrial network (stained red). The nucleus of the cell is stained in blue.

CHAPTER CONCEPTS

- Extranuclear inheritance occurs when phenotypes result from genetic influence other than the biparental transmission of genes located on chromosomes housed within the nucleus.
- Organelle heredity, an example of extranuclear inheritance, is due to the transmission of genetic information contained in mitochondria or chloroplasts, most often from only one parent.
- Traits determined by mitochondrial DNA are most often transmitted uniparentally through the maternal gamete, while traits determined by chloroplast DNA may be transmitted uniparentally or biparentally.
- Mitochondrial mutations have been linked to many human disease conditions as well as to the aging process.
- Maternal effect, the expression of the maternal nuclear genotype during gametogenesis and during early development, may have a strong influence on the phenotype of an organism.

Throughout the history of genetics, occasional reports have challenged the basic tenet of Mendelian transmission genetics—that the phenotype is transmitted by nuclear genes located on the chromosomes of both parents. Observations have revealed inheritance patterns that fail to reflect Mendelian principles, and some indicated an apparent extranuclear influence on the phenotype. Before the role of DNA in genetics was firmly established, such observations were commonly regarded with skepticism. However, with the discovery of DNA in mitochondria and the increasing knowledge of molecular genetics, extranuclear inheritance was soon recognized as an important aspect of genetics.

There are several varieties of extranuclear inheritance. One major type is **organelle heredity.** In this type of inheritance, DNA contained in mitochondria or chloroplasts determines certain phenotypic characteristics of the offspring. Examples are often recognized on the basis of the uniparental transmission of these organelles, usually from the female parent through the egg to progeny. A second type, called **infectious heredity,** results from a symbiotic or parasitic association with a microorganism. In such cases, an inherited phenotype is affected by the presence of the microorganism in the cytoplasm of the host cells. A third variety involves the **maternal effect** on the phenotype, whereby nuclear gene products are stored in the egg and then transmitted through the ooplasm to offspring. These gene products are distributed to cells of the developing embryo and influence its phenotype.

The common element in all of these examples is the transmission of genetic information to offspring through the cytoplasm rather than through the nucleus, most often from only one of the parents. This shall constitute our definition of **extranuclear inheritance**.

9.1 Organelle Heredity Involves DNA in Chloroplasts and Mitochondria

In this section and in Section 9.2, we will examine examples of inheritance patterns arising from chloroplast and mitochondrial function. Such patterns are now appropriately called *organelle heredity***.** While the exact mechanisms of transmission of traits were not initially clear, their inheritance appeared to be linked to something in the cytoplasm rather than to genes in the nucleus. Often (but not in all cases), the traits appeared to be transmitted from the maternal parent through the ooplasm, causing the results of reciprocal crosses to vary.

Analysis of the inheritance patterns resulting from mutant alleles in chloroplasts and mitochondria has been difficult for two major reasons. First, the function of these organelles is dependent on gene products from both nuclear and organelle DNA, making the discovery of the genetic origin of mutations affecting organelle function difficult. Second, large numbers of these organelles are contributed to each progeny cell following cell division. If only one or a few of the organelles acquire a new mutation or contain an existing one in a cell with a population of mostly normal organelles, the corresponding mutant phenotype may not be revealed, since the organelles lacking the mutation perform the wild-type function for the cell. Such variation in the genetic content of organelles is called **heteroplasmy.** Analysis is thus much more complex for traits controlled by genes encoded by organelle DNA than for Mendelian characters controlled by nuclear genes.

We will begin our discussion with several of the classical examples that ultimately called attention to organelle heredity. After that, we will discuss information concerning the DNA and the resultant genetic function in each organelle.

Chloroplasts: Variegation in Four O'Clock Plants

In 1908, Carl Correns (one of the rediscoverers of Mendel's work) provided the earliest example of inheritance linked to chloroplast transmission. Correns discovered a variant of the four o'clock plant, *Mirabilis jalapa,* in which some branches had white leaves, some had green, and some had variegated leaves. The completely white leaves and the white areas in variegated leaves lack chlorophyll that otherwise provides green color. Chlorophyll is the light-absorbing pigment made within chloroplasts.

Correns was curious about how inheritance of this phenotypic trait occurred. Inheritance in all possible combinations of crosses is strictly determined by the phenotype of the ovule source (**Figure 9.1**). For example, if the seeds (representing the progeny) were derived from ovules on branches with green leaves, all progeny plants bore only green leaves, regardless of the phenotype of the source of pollen. Correns concluded that inheritance was transmitted through the cytoplasm of the maternal parent because the pollen, which contributes little or no cytoplasm to the zygote, had no apparent influence on the progeny phenotypes.

Since leaf coloration is a function of the chloroplast, genetic information either contained in that organelle or somehow present in the cytoplasm and influencing the chloroplast must be responsible for the inheritance pattern. It now seems certain that the genetic "defect" that eliminates the green chlorophyll in the white patches on leaves is a mutation in the DNA housed in the chloroplast.

Chloroplast Mutations in *Chlamydomonas*

The unicellular green alga *Chlamydomonas reinhardtii* has provided an excellent system for the investigation of plastid inheritance. This haploid eukaryotic organism has a single

	Location of Ovule		
Source of Pollen	White branch	Green branch	Variegated branch
White branch	White	Green	White, green, or variegated
Green branch	White	Green	White, green, or variegated
Variegated branch	White	Green	White, green, or variegated

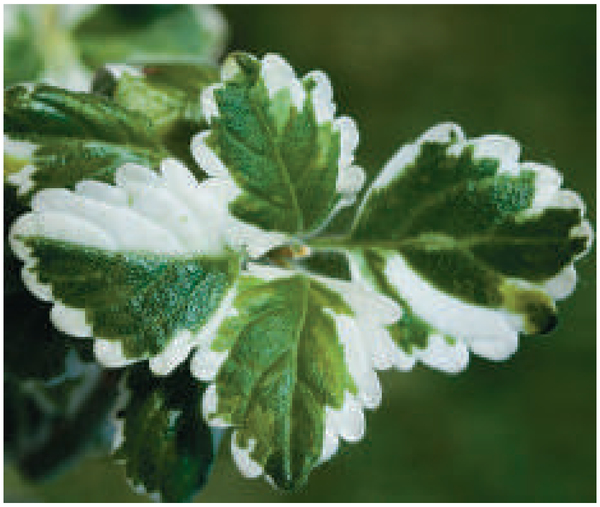
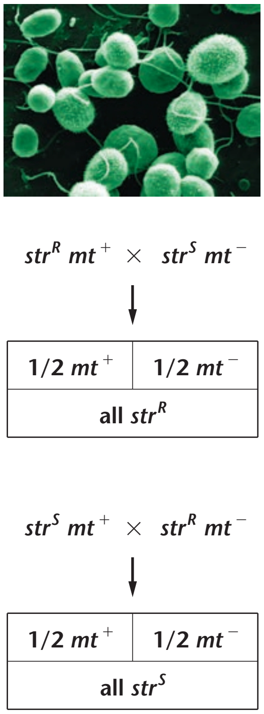


FIGURE 9.1 Offspring of crosses involving leaves from various branches of variegated four o'clock plants. The photograph illustrates variegation in the leaves of the madagascar spur.

large chloroplast containing about 75 copies of a circular double-stranded DNA molecule. Matings that reestablish diploidy are immediately followed by meiosis, and the various stages of the life cycle are easily studied in culture in the laboratory. The first known cytoplasmic mutation, streptomycin resistance (*str*^{R}) in *Chlamydomonas,* was reported in 1954 by Ruth Sager. Although *Chlamydomonas*'s two mating types, *mt*$^+$ and *mt*$^-$*,* appear to make equal cytoplasmic contributions to the zygote, Sager determined that the *str*^{R} phenotype is transmitted only through the *mt*$^+$ parent (**Figure 9.2**). Reciprocal crosses between sensitive and resistant strains yield different results depending on the genotype of the *mt*$^+$ parent, which is expressed in all offspring. As shown in the figure, one-half of the offspring are *mt*$^+$ and one-half of them are *mt*$^-$, indicating that mating type is controlled by a nuclear gene that segregates in a Mendelian fashion.

Following fertilization, which involves the fusion of two cells of opposite mating type, the single chloroplasts of the two mating types fuse. After the resulting zygote has undergone meiosis and haploid cells are produced, it is apparent that the genetic information in the chloroplast of progeny cells is derived only from the *mt*$^+$ parent. The genetic information originally present within the *mt*$^-$ chloroplast has degenerated.

Since Sager's discovery, a number of other *Chlamydomonas* mutations (including resistance to, or dependence on, a variety of bacterial antibiotics) that show a similar uniparental inheritance pattern have been discovered. These mutations have all been linked to the transmission of the chloroplast, and their study has extended our knowledge of chloroplast inheritance.



$str^R\ mt^+ \times str^S\ mt^-$
1/2 mt^+
1/2 mt^-
all str^R
$str^S\ mt^+ \times str^R\ mt^-$
1/2 mt^+
1/2 mt^-
all str^S

FIGURE 9.2 The results of reciprocal crosses between streptomycin-resistant (*str*^{R}) and streptomycin-sensitive (*str*^{S}) strains in the green alga *Chlamydomonas* (shown in the photograph).

NOW SOLVE THIS

9.1 *Chlamydomonas*, a eukaryotic green alga, may be sensitive to the antibiotic erythromycin, which inhibits protein synthesis in bacteria. There are two mating types in this alga, *mt*$^+$ and *mt*$^-$. If an *mt*$^+$ cell sensitive to the antibiotic is crossed with an *mt*$^-$ cell that is resistant, all progeny cells are sensitive. The reciprocal cross (*mt*$^+$ resistant and *mt*$^-$ sensitive) yields all resistant progeny cells. Assuming that the mutation for resistance is in the chloroplast DNA, what can you conclude from the results of these crosses?

- **HINT:** *This problem involves an understanding of the cytoplasmic transmission of organelles in unicellular algae. The key to its solution is to consider the results you would expect from two possibilities: that inheritance of the trait is uniparental or that inheritance is biparental.*

The inheritance of phenotypes influenced by mitochondria is also uniparental in *Chlamydomonas*. However, studies of the transmission of several cases of antibiotic resistance governed by mitochondria have shown that it is the *mt*$^-$ parent that transmits the mitochondrial genetic information to progeny cells. This is just the opposite of what occurs with chloroplast-derived phenotypes, such as *str*^{R}. The significance of inheriting one organelle from one parent and the other organelle from the other parent is not yet established.

Mitochondrial Mutations: Early Studies in *Neurospora* and Yeast

As alluded to earlier, mutations affecting mitochondrial function have also been discovered and studied, revealing that mitochondria, too, contain a distinctive genetic system. As with chloroplasts, mitochondrial mutations are transmitted through the cytoplasm during reproduction. In 1952, Mary B. Mitchell and Herschel K. Mitchell studied the pink bread mold *Neurospora crassa* (**Figure 9.3**). They discovered a slow-growing mutant strain and named it *poky*. (It is also designated *mi-1,* for *maternal inheritance*.) Slow growth is associated with impaired mitochondrial function, specifically caused by the absence of several cytochrome proteins essential for electron transport. In the absence of cytochromes, aerobic respiration leading to ATP synthesis is curtailed. Results of genetic crosses between wild-type and *poky* strains suggest that the trait is maternally inherited. If one mating type is *poky* and the other is wild type, all progeny colonies are *poky,* yet the reciprocal cross produces normal wild-type colonies.

FIGURE 9.3 The bread mold *Neurospora crassa,* grown in a Petri dish.

Another study of mitochondrial mutations has been performed with the yeast *Saccharomyces cerevisiae*. The first such mutation, described by Boris Ephrussi and his coworkers in 1956, was named *petite* because of the small size of the yeast colonies (**Figure 9.4**). Many independent *petite* mutations have since been discovered and studied, and all have a common characteristic: a deficiency in cellular respiration involving abnormal electron transport, as performed by mitochondria. This organism is a *facultative anaerobe* (an organism that can function both with and without the presence of oxygen), so in the absence of oxygen it can grow by fermenting glucose through glycolysis. Thus, it may survive the loss of mitochondrial function by generating energy anaerobically.

The complex genetics of *petite* mutations is diagrammed in **Figure 9.5**. A small proportion of these mutants are the result of nuclear mutations in genes whose products are transported to and function in mitochondria. They exhibit Mendelian inheritance and are thus called ***segregational petites.*** The remaining mutants demonstrate cytoplasmic transmission, indicating alterations in the DNA of the mitochondria. They produce one of two effects in matings. ***Neutral petites,*** when crossed to wild type, yield meiotic products (called ascospores) that give rise only to wild-type, or normal, colonies. The same pattern continues if the progeny of such crosses are backcrossed to *neutral petites*. The majority of "neutrals" lack mtDNA completely or have lost a substantial portion of it, so for their offspring to be normal, the neutrals must also be inheriting mitochondria capable of aerobic respiration from the normal parent following reproduction. This establishes that in yeast, mitochondria are inherited from both parental cells. The functional mitochondria from the normal parent are replicated in offspring and support aerobic respiration.

The third mutational type, ***suppressive petites,*** provides different results. Crosses between mutant and wild type give rise to diploid zygotes, which after meiosis, yield haploid cells that all express the *petite* phenotype. Assuming that the offspring have received mitochondria from both parents, the *petite* cells behave as what is called a **dominant-negative mutation,** which somehow suppresses the function of the wild-type mitochondria.

Two major hypotheses concerning the organelle DNA have been advanced to explain this suppressiveness. One explanation suggests that the mutant (or deleted) DNA in the mitochondria (mtDNA) replicates more rapidly, resulting in the mutant mitochondria "taking over" or dominating the phenotype by numbers alone. The second

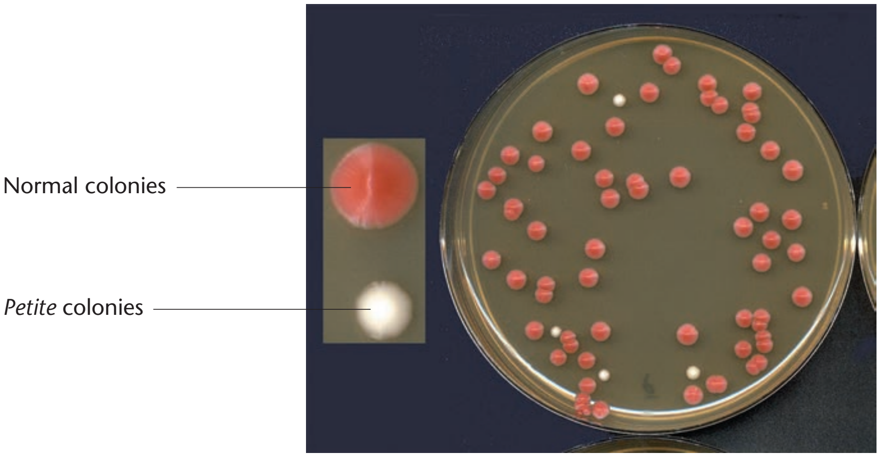


FIGURE 9.4 A comparison of normal versus *petite* colonies in the yeast *Saccharomyces cerevisiae*. The larger normal colonies appear pink, while the smaller mutant *petite* colonies appear white..

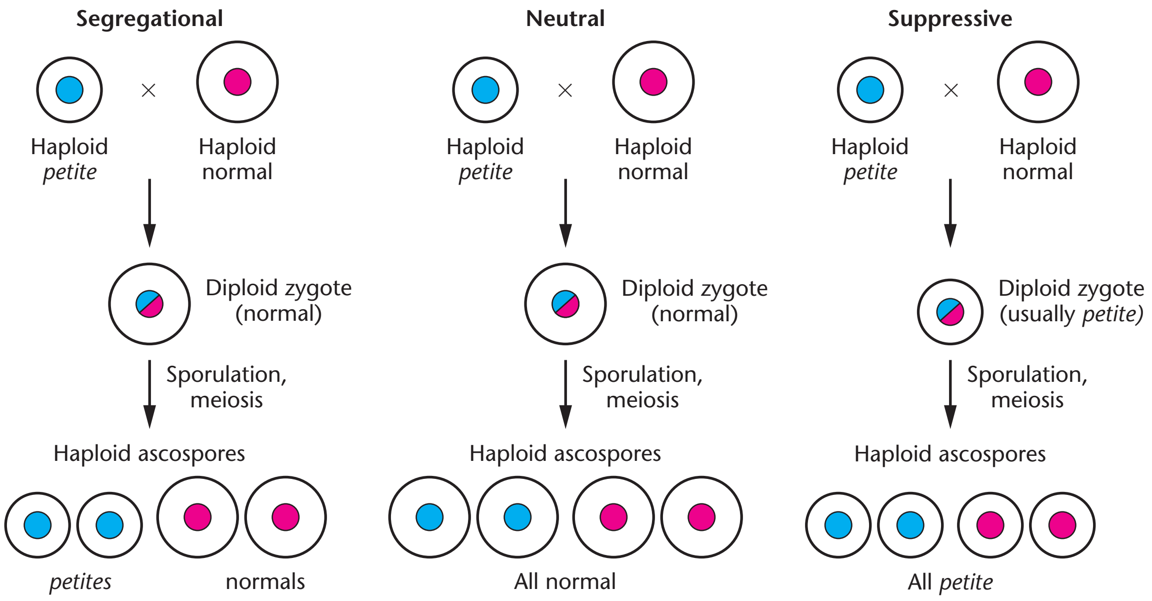


FIGURE 9.5 The outcome of crosses involving the three types of *petite* mutations affecting mitochondrial function in the yeast *Saccharomyces cerevisiae*.

explanation suggests that recombination occurs between the mutant and wild-type mtDNA, introducing errors into or disrupting the normal mtDNA. It is not yet clear which one, if either, of these explanations is correct.

NOW SOLVE THIS

9.2 In aerobically cultured yeast, a *petite* mutant is isolated. To determine the type of mutation causing this phenotype, the *petite* and wild-type strains are crossed. Such a cross has three potential outcomes.

(a) all wild type
(b) some *petite*, some wild type
(c) all *petite*

For each set of results, what conclusion about the type of *petite* mutation is justified?

■ **HINT:** *This problem involves the understanding that the* petite *phenotype is related to mitochondrial function and to how mitochondria are inherited. The key to its solution is to remember that in yeast, inheritance of mitochondria is biparental.*

9.2 Knowledge of Mitochondrial and Chloroplast DNA Helps Explain Organelle Heredity

That both mitochondria and chloroplasts contain their own DNA and a system for expressing genetic information was first suggested by the discovery of mutations and the resultant inheritance patterns in plants, yeast, and other fungi, as already discussed. Because both mitochondria and chloroplasts are inherited through the maternal cytoplasm in most organisms, and because mutations could be linked hypothetically to the altered function of either chloroplasts or mitochondria, geneticists set out to look for more direct evidence of DNA in these organelles. Not only was unique DNA found to be a normal component of both mitochondria and chloroplasts, but careful examination of the nature of this genetic information would provide essential clues as to the evolutionary origin of these organelles.

Organelle DNA and the Endosymbiotic Theory

Electron microscopists not only documented the presence of DNA in mitochondria and chloroplasts, but they also saw that it exists there in a form quite unlike the form seen in the nucleus of the eukaryotic cells that house these organelles (Figures 9.6 and 9.7). The DNA in chloroplasts and mitochondria looks remarkably similar to the DNA seen in bacteria. This similarity, along with the observation of the presence of a unique genetic system capable of organelle-specific transcription and translation, led Lynn Margulis and others to the postulate known as the **endosymbiotic theory.** Basically, the theory states that mitochondria and chloroplasts arose independently about 2 billion years ago from free-living protobacteria (primitive bacteria). Progenitors possessed the abilities now attributed to these organelles—aerobic respiration and photosynthesis, respectively. This idea proposes that these ancient bacteria-like cells were engulfed by larger primitive eukaryotic cells, which originally lacked the ability to respire aerobically or to capture energy from sunlight. A beneficial, symbiotic

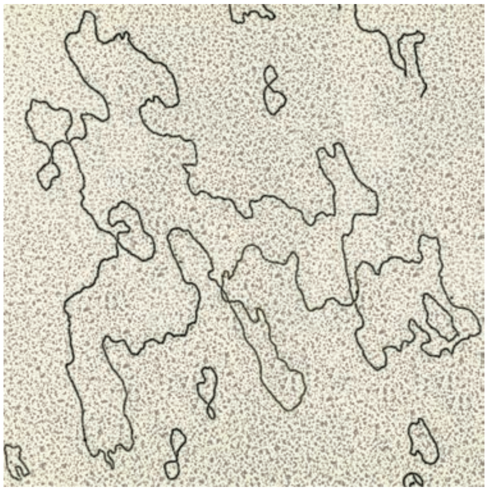


FIGURE 9.6 An electron micrograph of chloroplast DNA (cpDNA) derived from lettuce. Genes encoded by cpDNA impart photosynthetic function to the plant as well as providing the basis for transcription and translation of genetic information within the organelle.

relationship subsequently developed, whereby the bacteria eventually lost their ability to function autonomously, while the eukaryotic host cells gained the ability to perform either oxidative respiration or photosynthesis, as the case may be. Although some questions remain unanswered, evidence continues to accumulate in support of this theory, and its basic tenets are now widely accepted.

A brief examination of modern-day mitochondria will help us better understand endosymbiotic theory. During the course of evolution subsequent to the invasion event, distinct branches of diverse eukaryotic organisms arose. As the evolution of the host cells progressed, the companion bacteria also underwent their own independent changes. The primary alteration was the transfer of many of the genes from the invading bacterium to the nucleus of the host. The *products* of these genes, though still functioning in the organelle, are nevertheless now encoded and transcribed in the nucleus and translated in the cytoplasm prior to their transport into the organelle. The amount of DNA remaining today in the typical mitochondrial genome is minuscule compared with that in the free-living bacteria from which it was derived. The most gene-rich organelles now have fewer than 10 percent of the genes present in the smallest bacterium known.

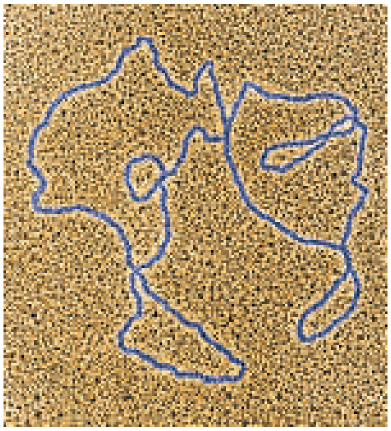


FIGURE 9.7 Electron micrograph of mitochondrial DNA derived from the frog *Xenopus laevis*.

Similar changes have characterized the evolution of chloroplasts. In the subsequent sections of this chapter, we will explore in some detail what is known about modern-day chloroplasts and mitochondria.

Molecular Organization and Gene Products of Chloroplast DNA

The chloroplast, responsible for photosynthesis, contains both DNA (as a source of genetic information) and a complete protein-synthesizing apparatus. The molecular components of the chloroplast's translation apparatus are derived from both nuclear and organelle genetic information.

Chloroplast DNA (cpDNA), shown in Figure 9.6, is fairly uniform in size among different organisms, ranging between 100 and 225 kb in length. It shares many similarities to DNA found in bacterial cells. It is circular and double stranded, and it is free of the associated proteins characteristic of eukaryotic DNA. Compared with nuclear DNA from the same organism, it invariably shows a different density and base composition.

The size of cpDNA is much larger than that of mtDNA. To some extent, this can be accounted for by a larger number of genes. However, most of the difference appears to be due to the presence in cpDNA of many long noncoding nucleotide sequences both between and within genes, the latter being introns, or noncoding DNA characteristic of eukaryotic nuclear DNA (see Chapter 13). Duplications of many DNA sequences are also present. Since such noncoding sequences vary in different plants, they indicate that independent evolution occurred in chloroplasts following their initial invasion of a primitive eukaryotic-like cell.

In the green alga *Chlamydomonas,* there are about 75 copies of the chloroplast DNA molecule per organelle. In higher plants, such as the sweet pea, multiple copies of the DNA molecule are also present in each organelle, but the molecule (134 kb) is considerably smaller than that in *Chlamydomonas* (195 kb). Interestingly, genetic recombination between the multiple copies of DNA within chloroplasts has been documented in some organisms.

Numerous gene products encoded by chloroplast DNA function during translation within the organelle. Two sets each of the genes coding for the ribosomal RNAs—16*S,* and 23*S* rRNA—are present. *S* refers to the Svedberg coefficient (described in Chapter 10), which is related to the molecule's size and shape. In addition, cpDNA encodes numerous transfer RNAs (tRNAs), as well as many ribosomal proteins specific to the chloroplast ribosomes. In the liverwort, whose cpDNA was the first to be sequenced, there are genes encoding 30 tRNAs, RNA polymerase, multiple rRNAs, and numerous ribosomal proteins. The variations in the gene products encoded in the cpDNA of different plants again attest to the independent evolution that occurred within chloroplasts.

Chloroplast ribosomes differ significantly from those present in the cytoplasm and encoded by nuclear genes. They have a Svedberg coefficient slightly less than 70*S*, which characterizes bacterial ribosomes. Even though some chloroplast ribosomal proteins are encoded by chloroplast DNA and some by nuclear DNA, most, if not all, such proteins are chemically distinct from their counterparts present in cytoplasmic ribosomes. Both observations provide direct support for the endosymbiotic theory.

Still other chloroplast genes specific to the photosynthetic function, the major role normally associated with this organelle, have been identified. For example, in the moss, there are 92 chloroplast genes encoding proteins that are part of the thylakoid membrane, a cellular component integral to the light-dependent reactions of photosynthesis. Mutations in these genes may inactivate photosynthesis. A typical distribution of genes between the nucleus and the chloroplast is illustrated by one of the major photosynthetic enzymes, ribulose-1-5-bisphosphate carboxylase (known as *Rubisco*). This enzyme has its small subunit encoded by a nuclear gene, whereas the large subunit is encoded by cpDNA.

Molecular Organization and Gene Products of Mitochondrial DNA

Extensive information is also available concerning the structure and gene products of mitochondrial DNA (mtDNA). In most eukaryotes, mtDNA exists as a double-stranded, closed circle (Figure 9.7) that, like cpDNA, is free of the chromosomal proteins characteristic of eukaryotic chromosomal DNA. An exception is found in some ciliated protozoans, in which the DNA is linear.

In size, mtDNA is much smaller than cpDNA and varies greatly among organisms, as demonstrated in **Table 9.1**. In a variety of animals, including humans, mtDNA consists of about 16,000 to 18,000 bp (16 to 18 kb). However, yeast (*Saccharomyces*) mtDNA consists of 75 kb. Plants typically exceed this amount—367 kb is present in mitochondria in the mustard plant, *Arabidopsis*. Vertebrates have 5 to 10 such DNA molecules per organelle, while plants have 20 to 40 copies per organelle. The number of genes in the mtDNA genome is also variable, from less than 12 genes to 100 genes, depending on the species.

TABLE 9.1 The Size of mtDNA in Different Organisms

Organism	Size (kb)
Homo sapiens (human)	16.6
Mus musculus (mouse)	16.2
Xenopus laevis (frog)	18.4
Drosophila melanogaster (fruit fly)	18.4
Saccharomyces cerevisiae (yeast)	75.0
Pisum sativum (pea)	110.0
Arabidopsis thaliana (mustard plant)	367.0

There are several other noteworthy aspects of mtDNA. With only rare exceptions, and unlike cpDNA, introns are absent from mitochondrial genes, and gene repetitions are seldom present. Nor is there usually much in the way of intergenic spacer DNA. This is particularly true in species whose mtDNA is fairly small in size, such as humans. In *Saccharomyces,* with a much larger mtDNA molecule, much of the excess DNA is accounted for by introns and intergenic spacer DNA. As will be discussed in Chapter 13, the expression of mitochondrial genes uses several modifications of the otherwise standard genetic code. Also of interest is the fact that replication in mitochondria is dependent on enzymes encoded by nuclear DNA.

Human mtDNA encodes two ribosomal RNAs (rRNAs), 22 transfer RNAs (tRNAs), and 13 polypeptides essential to the oxidative respiration functions of the organelle. For instance, mitochondrial-encoded gene products are present in all of the protein complexes of the electron transport chain found in the inner membrane of mitochondria. In most cases, these polypeptides are part of multichain proteins, many of which also contain subunits that are encoded in the nucleus, synthesized in the cytoplasm, and then transported into the organelle. Thus, the protein-synthesizing apparatus and the molecular components for cellular respiration are jointly derived from nuclear and mitochondrial genes.

Like chloroplasts, mitochondria also have unique ribosomes involved in translation of genes encoded by mtDNA. Mitochondrial ribosomes of different species vary considerably in their Svedberg coefficients, ranging from 55*S* to 80*S,* while cytoplasmic ribosomes are uniformly 80*S*. The majority of proteins that function in mitochondria are encoded by nuclear genes. In fact, approximately 1500 nuclear-coded gene products are essential to biological activity in the organelle. They include, for example, DNA and RNA polymerases, initiation and elongation factors essential for translation, ribosomal proteins, aminoacyl tRNA synthetases, and several tRNA species. mRNA processing and posttranslational modifications of mitochondrial proteins also involve nuclear gene products. Notably, these imported components are distinct from their cytoplasmic counterparts, even though both sets are coded by nuclear genes, providing further support for the endosymbiotic theory. For example, the synthetase enzymes essential for charging mitochondrial tRNA molecules (a process essential to translation) show a distinct affinity for the mitochondrial tRNA species as compared with the cytoplasmic tRNAs. Similar affinity has been shown for the initiation and elongation factors. Furthermore, while bacterial and nuclear RNA polymerases are known to be composed of numerous subunits, the mitochondrial variety consists of only one polypeptide chain. This polymerase is generally sensitive to antibiotics that inhibit bacterial RNA synthesis, but not to eukaryotic inhibitors. The various contributions of some of the nuclear and mitochondrial gene products are contrasted in **Figure 9.8**.

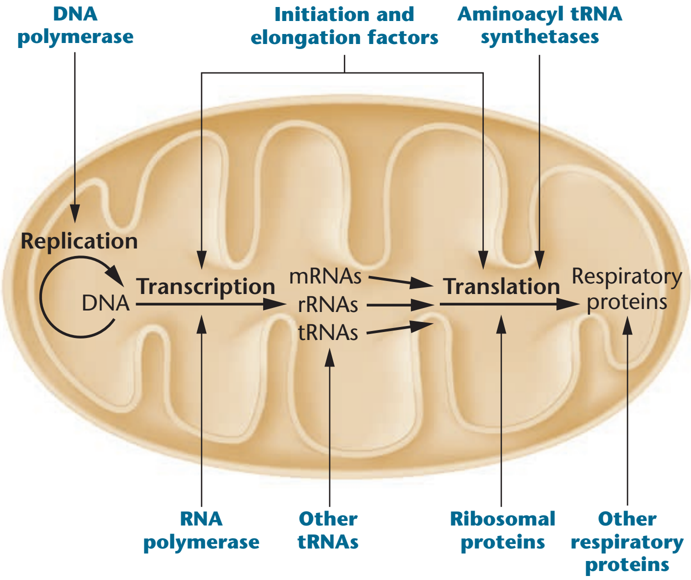


FIGURE 9.8 Gene products that are essential to mitochondrial function. Those shown entering the organelle are derived from the cytoplasm and encoded by nuclear genes.

NOW SOLVE THIS

9.3 DNA in human mitochondria encodes 22 different tRNA molecules. However, 32 different tRNA molecules are required for translation of proteins within mitochondria. Explain.

■ **HINT:** *This problem involves understanding the origin of mitochondria in eukaryotes. The key to its solution is to consider the endosymbiotic theory and its ramifications.*

For more practice, see Problems 17–19.

9.3 Mutations in Mitochondrial DNA Cause Human Disorders

The human mtDNA genome has been sequenced and contains 16,569 base pairs. While estimates suggest that the protobacteria thought to have given rise to mitochondria contained several thousand genes, human mtDNA has just 37 genes, all considered essential. Some ancestral genes are known to have been transferred to the nuclear genome, but large numbers of nonessential genes likely disappeared from the mitochondrial genome. Gene products encoded by mtDNA include 13 proteins required for aerobic cellular respiration. Because a cell's energy supply is largely dependent on aerobic cellular respiration to generate ATP, disruption of any mitochondrial gene by mutation may potentially have a severe impact on that organism. We have seen this in our previous discussion of the *petite* mutations in yeast, which would be lethal were it not for this organism's ability to respire anaerobically. In fact, mtDNA is particularly vulnerable to mutations, for at least three reasons. First, mtDNA does not have the structural protection from mutations provided by histone proteins present in nuclear DNA. Second, DNA repair mechanisms for mtDNA are limited. Third, mitochondria concentrate highly mutagenic **reactive oxygen species (ROS)** generated by cell respiration. In such a confined space, ROS are toxic to the contents of the organelle and are known to damage proteins, lipids, and mtDNA. Ultimately, this increases the frequency of point mutations and deletions in the mitochondria. As a result, mutations in mtDNA occur at a rate that is at least tenfold higher than in nuclear DNA.

Individual cells can have several thousand mitochondria. Fortunately, a zygote receives a large number of organelles through the egg, so if only one organelle or a few of them contain a mutation, its impact is greatly diluted by the many mitochondria that lack the mutation and function normally. During early development, cell division disperses the initial population of mitochondria present in the zygote, and in the newly formed cells, these organelles reproduce autonomously. Therefore, an embryo, and subsequently, the adult organism will have cells with a variable mixture of both normal and abnormal organelles. This variation in the genetic content of organelles, as indicated earlier in the chapter, is called *heteroplasmy*.

In order for a human disorder to be attributable to genetically altered mitochondria, several criteria must be met:

1. Inheritance must exhibit a maternal rather than a Mendelian pattern.

2. The disorder must reflect a deficiency in the bioenergetic function of the organelle.
3. There must be a mutation in one or more of the mitochondrial genes.

Today over 150 highly diverse human syndromes are known to demonstrate these characteristics. Some are expressed in infancy, while others are not expressed until adulthood. Some impact multiple organ systems, and some are highly organ-specific. For example, **myoclonic epilepsy and ragged-red fiber disease (MERRF)** demonstrates a pattern of inheritance consistent with maternal transmission. Only the offspring of affected mothers inherit the disorder; the offspring of affected fathers are normal. Individuals with this rare disorder express ataxia (lack of muscular coordination), deafness, dementia, and epileptic seizures. The disease is so named because of the presence of "ragged-red" skeletal muscle fibers that exhibit blotchy red patches resulting from the proliferation of aberrant mitochondria (**Figure 9.9**). Brain function, which has a high energy demand, is affected in this disorder, leading to the neurological symptoms described.

Analysis of mtDNA from patients with MERRF has revealed a mutation in one of the 22 mitochondrial genes encoding a transfer RNA. Specifically, the gene encoding tRNA^{Lys} (the tRNA that delivers lysine during translation) contains an A-to-G transition within its sequence. This genetic alteration apparently interferes with the capacity for translation within the organelle, which in turn leads to the various manifestations of the disorder.

The cells of affected individuals exhibit heteroplasmy, containing a mixture of normal and abnormal mitochondria. Different patients contain different proportions of the two, and even different cells from the same patient exhibit various levels of abnormal mitochondria. Were it not for heteroplasmy, the mutation would very likely be lethal, testifying to the essential nature of mitochondrial function and its reliance on the genes encoded by mtDNA within the organelle.

A second disorder, **Leber's hereditary optic neuropathy (LHON),** also exhibits maternal inheritance as well as mtDNA lesions. The disorder is characterized by sudden bilateral blindness. The average age of vision loss is 27, but onset is quite variable. Four mutations have been identified, all of which disrupt normal oxidative phosphorylation, the final pathway of respiration in cells. More than 50 percent of cases are due to a mutation at a specific position in the mitochondrial gene encoding a subunit of NADH dehydrogenase. This mutation is transmitted maternally through the mitochondria to all offspring. Noteworthy is the observation that in many instances of LHON, there is no family history; a significant number of cases are "sporadic," resulting from newly arisen mutations.

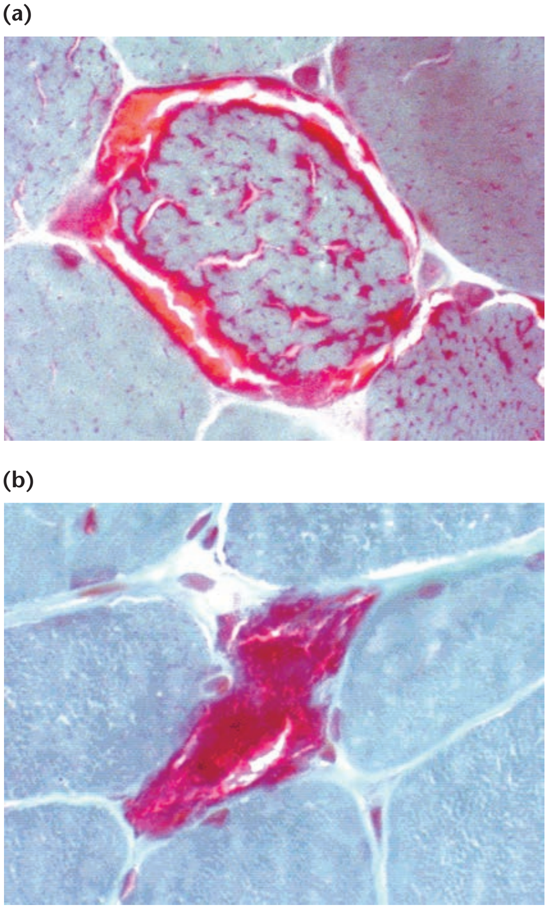


FIGURE 9.9 Ragged-red fibers in skeletal muscle cells from patients with the mitochondrial disease MERRF. (a) The muscle fiber has mild proliferation of mitochondria. (See red rim and speckled cytoplasm.) (b) Marked proliferation in which mitochondria have replaced most cellular structures.

Individuals severely affected by a third disorder, **Kearns-Sayre syndrome (KSS),** lose their vision, experience hearing loss, and display heart conditions. The genetic basis of KSS involves deletions at various positions within mtDNA. Many KSS patients are symptom-free as children but display progressive symptoms as adults. The proportion of mtDNAs that reveal deletion mutations increases as the severity of symptoms increases.

Mitochondria, Human Health, and Aging

The study of hereditary mitochondrial-based disorders provides insights into the critical importance of this organelle during normal development. In fact, mitochondrial dysfunction seems to be implicated in most all major human disease conditions, including anemia, blindness, Type II (late-onset) diabetes, autism, atherosclerosis, infertility, neurodegenerative diseases such as Parkinson, Alzheimer, and Huntington disease, schizophrenia and bipolar disorders, and a variety of cancers. It is becoming evident, for example, that mutations in mtDNA are present in such human malignancies as skin, colorectal, liver, breast, pancreatic, lung, prostate, and bladder cancers.

Over 400 mtDNA mutations associated with more than 150 distinct mtDNA-based genetic syndromes have been identified. Genetic tests for detecting mutations in the mtDNA genome that may serve as early-stage disease markers have been developed. For example, mtDNA mutations in skin cells have been detected as a biomarker of cumulative exposure of ultraviolet light and development of skin cancer. However, it is still unclear whether mtDNA mutations are causative effects contributing to development of malignant tumors or whether they are the consequences of tumor formation. Nonetheless, there is an interesting link between mtDNA mutations and cancer, including data suggesting that many chemical carcinogens have significant mutation effects on mtDNA.

The study of hereditary mitochondrial-based disorders has also suggested a link between the progressive decline of mitochondrial function and the aging process. It has been hypothesized that the accumulation of sporadic mutations in mtDNA leads to an increased prevalence of defective mitochondria (and the concomitant decrease in the supply of ATP) in cells over a lifetime. This condition in turn plays a significant role in aging. It has been suggested that cells require a threshold level of ATP production resulting from oxidative phosphorylation for normal function. When the level drops below this threshold, the aging process is accelerated.

Many studies have now documented that aging tissues contain mitochondria with increased levels of DNA damage. The major question is whether such changes are simply biomarkers of the aging process or whether they lead to a decline in physiological function, which in turn, contributes significantly to aging. In support of the latter hypothesis, one study links age-related muscle fiber atrophy in rats to deletions in mtDNA and electron transport abnormalities. Such deletions appear to be present in the mitochondria of atrophied muscle fibers, but are absent from fibers in regions of normal tissue.

It is important to note that mutations in the nuclear genome also impact mitochondrial function and human disease and aging. For example, another study involving genetically altered mice is most revealing. Such mice have a nuclear gene altered that diminishes proofreading during the replication of mtDNA. These mice display reduced fertility and accumulate mutations over time at a much higher rate than is normal. These mice also show many characteristics of premature aging, as observed by loss and graying of hair, reduction in bone density and muscle mass, decline in fertility, anemia, and reduced life span.

Other recent studies have demonstrated that many nuclear gene products, as many as 75, function during aerobic respiration within mitochondria by binding to proteins encoded by mtDNA. These, and other studies, continue to speak to the importance of generating adequate ATP as a result of oxidative phosphorylation under the direction of DNA in mitochondria. As cells undergo genetic damage, which appears to be a natural phenomenon, their function declines, which may be an underlying factor in aging as well as in the progression of age-related disorders.

Future Prevention of the Transmission of mtDNA-Based Disorders

Estimates suggest that 1 in 5000 humans either have a mtDNA-based disease or are at risk for developing such a genetic disorder. Many mtDNA-based genetic disorders can now be detected by genetic testing, but there are currently no cures for these disorders. However, new therapeutic approaches are being developed that can prevent *transmission of mtDNA mutations* from an affected mother to her offspring. In 2010, scientists demonstrated that the mtDNA genome can be replaced in the eggs of a nonhuman primate, the rhesus monkey (*Macaca mulatta*). As illustrated in **Figure 9.10**, this was accomplished by transplanting the nuclear genome from an egg of one mother to an *enucleated* recipient egg (one with its nucleus removed) of another female with a normal content of mitochondria. These reconstructed eggs were fertilized and implanted in the uterus of the original mother. The result was the production of three-parent offspring containing nuclear DNA from the transplanted maternal nuclear genome of one female, mitochondria from the recipient egg of another female, and a paternal genome from the sperm donor. As these monkeys have progressed into adulthood, so far they appear to be normal.

Referred to as **mitochondrial replacement therapy (MRT)** or **three-parent *in vitro* fertilization,** this technology has now been extended to human reproductive research. Near the end of 2012, MRT was implemented with human eggs and *in vitro* fertilization. The successful zygotes developed normally to the blastocyst stage (the 100-cell stage), which was then used to develop cultured embryonic cell lines, which were tested in numerous ways. These cell lines behaved and tested similarly to those that had not undergone mitochondrial replacement

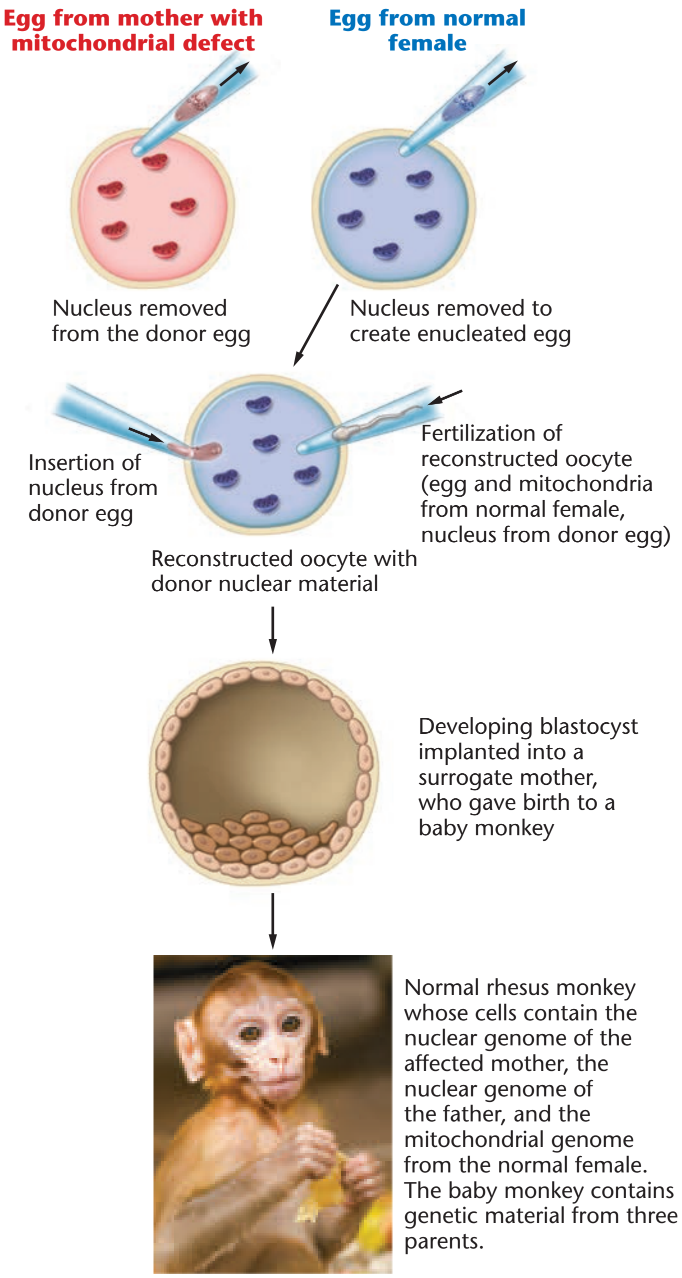

Egg from mother with mitochondrial defect
Egg from normal female
Nucleus removed from the donor egg
Nucleus removed to create enucleated egg
Insertion of nucleus from donor egg
Fertilization of reconstructed oocyte (egg and mitochondria from normal female, nucleus from donor egg)
Reconstructed oocyte with donor nuclear material
Developing blastocyst implanted into a surrogate mother, who gave birth to a baby monkey
Normal rhesus monkey whose cells contain the nuclear genome of the affected mother, the nuclear genome of the father, and the mitochondrial genome from the normal female. The baby monkey contains genetic material from three parents.

FIGURE 9.10 Illustration of mitochondrial swapping, working with oocytes from rhesus monkeys. The nucleus from one oocyte (red) with defective mitochondria was removed and transferred to an enucleated egg containing normal mitochondria. The father's sperm was injected and following fertilization, the developing blastocyst was implanted in a surrogate mother. A healthy monkey resulted. This mtDNA swapping approach has the potential to prevent transmission of mtDNA-based mutations if used for *in vitro* fertilization.

therapy. Thus, this technology may potentially offer a future treatment option for preventing mtDNA disease transmission in families with known histories of mtDNA-based disorders.

Several technical and ethical hurdles need to be overcome if MRT is to be used for humans. For instance, current techniques do not fully eliminate the transfer of some defective mitochondria into the donor egg. Whether enough defective mitochondria are transferred to impact the health of the embryo is not clear. Also, is it ethically acceptable to produce offspring with genetic contributions from three parents? (See the Genetics, Ethics, and Society feature on page 209.) In 2015, the United Kingdom became the first nation to legalize three-person *in vitro* fertilization techniques, designed to prevent children from inheriting diseases caused by mutations in mtDNA. Thus women with mtDNA-based diseases could possibly bear healthy children derived from their own nuclear DNA and mitochondria derived from a third party.

NOW SOLVE THIS

9.4 Given the maternal origin of mitochondria, from mother to offspring through the egg, and coupled with the relatively high mutation rate of mitochondrial DNA, one would expect that a high percentage of mutated, nonfunctional mitochondria would accumulate in cells of adults. One could envision that as a result, a mitochondrial meltdown would eventually occur as generations ensue. Such is not the case, however, and data indicate that over generations, mutations in mtDNA tend to be reduced in frequency in a given lineage. How might this phenomenon be explained?

- **HINT:** *This problem involves an understanding of heteroplasmy in mitochondria. The key to its solution is to consider factors that drive evolution, particularly natural selection, and to apply this concept to the presence of many mitochondria in each cell.*

9.4 In Maternal Effect, the Maternal Genotype Has a Strong Influence during Early Development

In **maternal effect,** also referred to as *maternal influence,* an offspring's phenotype for a particular trait is under the control of nuclear gene products present in the egg. This is in contrast to biparental inheritance, where both parents transmit information on genes in the nucleus that determines the offspring's phenotype. In cases of maternal effect, the nuclear genes of the female gamete are transcribed, and the genetic products (either proteins or untranslated RNAs) accumulate in the egg cytoplasm. After fertilization, these products are distributed among newly formed cells and influence the patterns or traits established during early development. Two examples will illustrate such an influence of the maternal genome on particular traits.

Lymnaea Coiling

Shell coiling (chirality) in the snail *Lymnaea peregra* is an excellent example of maternal effect on a permanent rather than a transitory phenotype. Some strains of this snail have left-handed, or sinistrally, coiled shells (*dd*), while others have right-handed, or dextrally, coiled shells (*DD* or *Dd*). These snails are hermaphroditic and may undergo either cross- or self-fertilization, providing a variety of types of matings.

Figure 9.11 illustrates the results of reciprocal crosses between true-breeding snails. As you can see, these crosses yield different outcomes, even though both are between

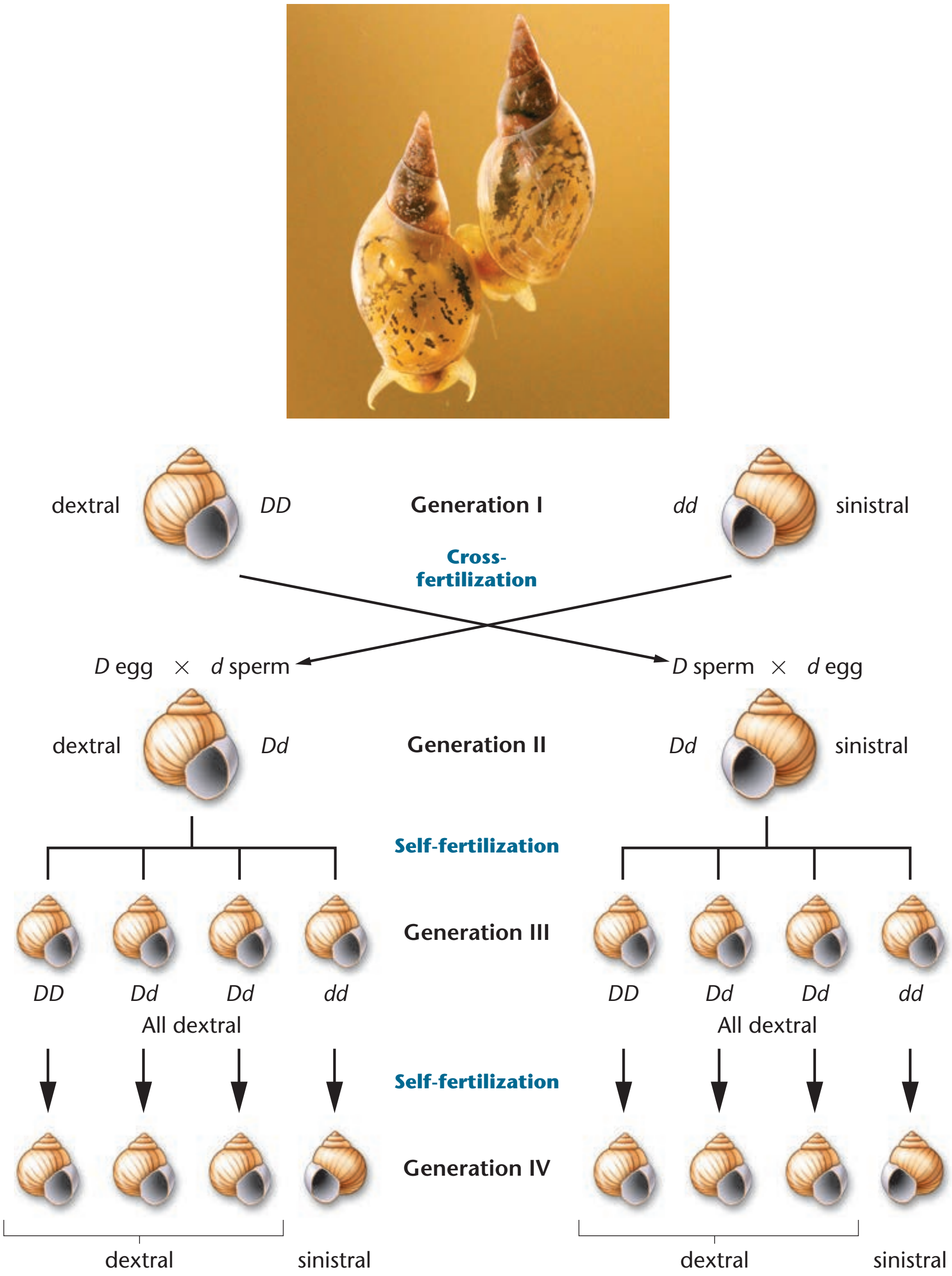


FIGURE 9.11 Inheritance of coiling in the snail *Lymnaea peregra*. Coiling is either dextral (right-handed) or sinistral (left-handed). A maternal effect is evident in generations II and III, where the genotype of the maternal parent, rather than the offspring's own genotype, controls the phenotype of the offspring. The photograph illustrates right- versus left-handed coiled snails.

sinistral and dextral organisms and produce all heterozygous offspring. Examination of the progeny reveals that their phenotypes depend on the genotype of the female parent. If we adopt that conclusion as a working hypothesis, we can test it by examining the offspring in subsequent generations of self-fertilization events. In each case, the hypothesis is upheld. Ovum donors that are *DD* or *Dd* produce only dextrally coiled progeny. Maternal parents that are *dd* produce only sinistrally coiled progeny. The coiling pattern of the progeny snails is determined by *the genotype of the parent producing the egg, regardless of the phenotype of that parent*.

Investigation of the developmental events in *Lymnaea* reveals that the orientation of the cleavage pattern following the eight-cell embryo stage determines the direction of coiling. Cell—cell interaction is influenced by a maternal gene in a way that establishes the orientation of dividing cells. The dextral allele (*D*) produces an active gene product that causes right-handed coiling. If ooplasm from dextral eggs is injected into uncleaved sinistral eggs, they cleave in a dextral pattern. However, in the converse experiment, sinistral ooplasm has no effect when injected into dextral eggs. Apparently, the sinistral allele is the result of a classic recessive mutation that encodes an inactive gene product.

We can conclude, then, that females that are either *DD* or *Dd* produce oocytes that synthesize the *D* gene product, which is stored in the ooplasm. Even if the oocyte contains only the *d* allele following meiosis and is fertilized by a *d*-bearing sperm, the resulting *dd* snail will be dextrally coiled (right handed).

Embryonic Development in *Drosophila*

A more recently documented example of maternal effect involves various genes that control embryonic development in *Drosophila melanogaster*. The genetic control of embryonic development in *Drosophila,* discussed in greater detail later in the text (see Chapter 23), is a fascinating story. The protein products of the maternal-effect genes function to activate other genes, which may in turn activate still other genes. This cascade of gene activity leads to a normal embryo whose subsequent development yields a normal adult fly. The extensive work by Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric Wieschaus (who shared the 1995 Nobel Prize for Physiology or Medicine for their findings) has clarified how these and other genes function. Genes that illustrate maternal effect have products that are synthesized by the developing egg and stored in the oocyte prior to fertilization. Following fertilization, these products create molecular gradients that determine spatial organization as development proceeds.

For example, the gene *bicoid* (*bcd*$^{+}$) plays an important role in specifying the development of the anterior portion of the fly. The RNA transcribed by this gene is deposited anteriorly in the egg (**Figure 9.12**), and upon translation, forms a gradient highest at the anterior end and gradually diluted posteriorly. Embryos derived from mothers who are homozygous for the mutant allele (*bcd*$^{-}$/*bcd*$^{-}$) fail to develop anterior areas that normally give rise to the head and thorax of the adult fly. Embryos whose mothers contain at least one wild-type allele (*bcd*$^{+}$) develop normally, even if the genotype of the embryo is homozygous for the mutation. Consistent with the concept of *maternal effect,* the *genotype of the female parent, not the genotype of the embryo* determines the phenotype of the offspring. Nüsslein-Volhard, and Wieschaus, using large-scale mutant screens, discovered many other maternal-effect genes critical to normal *Drosophila* development, influencing not only anterior and posterior morphogenesis, but also the expression of zygotic (nuclear) genes that control segmentation in this arthropod. When we return to our discussion of this general topic later in the text (see Chapter 23), we will see examples of other genes illustrating maternal effect that influence both anterior and posterior morphogenesis. Many of these genes function to regulate the spatial expression of "zygotic" (nuclear) genes that influence other aspects of development and that behave genetically in the conventional Mendelian fashion.

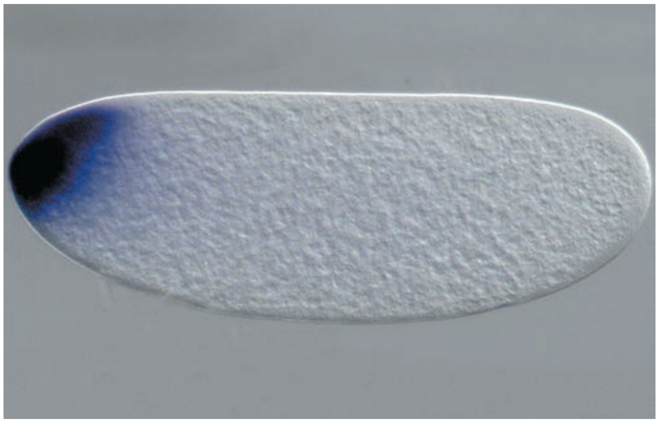


FIGURE 9.12 A gradient of *bicoid* mRNA (blue stain) as concentrated in the anterior region of the *Drosophila* embryo during early development.

GENETICS, ETHICS, AND SOCIETY

Mitochondrial Replacement and Three-Parent Babies

As many as 1 in 5000 people carry a potentially disease-causing mutation in most or all of their mitochondrial DNA (mtDNA). These mutations can lead to symptoms as varied as blindness, neurodegenerative defects, strokes, muscular dystrophies, and diabetes. In addition, mtDNA mutations can cause infertility, miscarriages, and the death of newborns and children. There are few treatments for mitochondrial diseases and, until recently, no potential cures.

In the 1990s, fertility clinics in the United States began using *cytoplasmic transfer* to treat some cases of infertility, including those caused by mitochondrial mutations. Cytoplasmic transfer involves the injection of cytoplasm containing normal mitochondria from a donor egg into a recipient egg prior to *in vitro* fertilization and implantation. Worldwide, more than 100 babies have been born following cytoplasmic transfer. In 2001, the U.S. Food and Drug Administration halted the procedure after two cases of X-chromosome abnormalities and one case of a developmental disorder appeared after the treatment. Since then, no research or treatments using cytoplasmic transfer have been permitted in the United States, although it is still used in other countries.

Since 2001, the United Kingdom and Japan have pioneered the use of other mitochondrial treatment methods. These methods are known generally as mitochondrial replacement therapies (MRTs). Presently, the two most commonly used MRTs are maternal spindle transfer (MST) and pronuclear transfer (PNT).

As described earlier in this chapter, MST is an *in vitro* fertilization method that involves the transfer of nuclear chromosomes at the metaphase II stage of meiosis from the patient's egg (containing defective mitochondria) into an enucleated donor egg (containing healthy mitochondria). The egg is then fertilized *in vitro*. PNT is a variation of MST. It involves *in vitro* fertilization of both the donor's egg and the patient's egg with sperm from the same donor. After fertilization, but before pronuclear fusion, the egg and sperm pronuclei from the patient's zygote are transferred to the donor's enucleated zygote. In both methods, the resulting hybrid zygotes are grown to the blastocyst stage, screened for genetic defects, and implanted into the patient. In both methods, the resulting offspring contain nuclear DNA from the mother and father, and mtDNA from the donor.

In 2015, the United Kingdom became the first country in the world to legalize the use of MRT methods for research in human embryos, and in 2016, it approved the limited use of MRT in humans. The worldwide media responded with headlines of "three-parent babies" and "children with three genetic parents," and the articles led to some confusion about the details of the techniques involved. Since then, the outburst of controversy has centered on the ethical and social implications of using MRT. Proponents of MRT welcome these mitochondrial treatments as significant contributions to the welfare of parents and children, who will no longer suffer from debilitating diseases. Opponents express reservations ranging from safety concerns to objections about human germ-line modification and damaging a child's identity.

Over the next decade, as more children are born with mitochondria from donor eggs, we will gain more perspective on the positive and negative aspects of MRT.

Your Turn

Take time, individually or in groups, to consider the following questions. Investigate the references dealing with the technical and ethical challenges surrounding mitochondrial replacement.

1. Summarize the ethical arguments used to support and oppose the use of MRT in humans. In your opinion, which arguments have validity, and why? How do the ethical arguments differ between PNT versus MST?

These topics are discussed in Gómez-Tatay, L. et al. (2017). Mitochondrial Modification Techniques and Ethical Issues. *J. Clin. Med*. 6, 25; DOI:10.3390/jcm6030025. *Also, see* Baylis, F. (2013). The Ethics of Creating Children with Three Genetic Parents. *Reproductive BioMed Online*, 26:531–534.

2. Much of the controversy surrounding MRT methods has been triggered by the phrase "three-parent babies" in media headlines. Do you think that this phrase is an accurate description of children born following mitochondrial replacement?

For a discussion of this question, see Dimond, R. (2015). Social and Ethical Issues in Mitochondrial Donation. *Br. Med. Bull*. 115:173–182.

CASE STUDY Is it all in the genes?

Marcia saw an ad on television for ancestry DNA testing and thought, "Why not?" She ordered a kit, swabbed her inner cheek, and returned the kit for analysis. Several weeks later, she was surprised to learn that she was 17 percent Native American. An elderly great aunt confirmed that her mother's family intermarried with members of Native American tribes in the Pacific Northwest in the early twentieth century. To investigate her maternal heritage, Marcia ordered a mitochondrial DNA (mtDNA) test, which confirmed her Native American ancestry. Based on these genetic results, she applied to several Native American tribes for enrollment as a tribal member. She was shocked when she was turned down. In discussions, tribal officials told her that DNA alone is not sufficient to define who is Native American. Tribal standards for enrollment vary, but usually cultural attributes such as knowledge of the language, customs, and history of the tribe are important considerations for enrollment decisions. Marcia was not satisfied and felt that she had a strong case based on biology alone. This series of events raises several questions:

1. Why did Marcia choose mitochondrial testing to determine her maternal heritage?
2. How many great-grandmothers does any individual (such as Marcia) have? How many of them contribute to the mitochondrial DNA that an individual (Marcia) carries?
3. How much importance should we place on the results of ancestral genetic testing especially when these results have social, political, and legal implications? Is it ethical to determine one's identity primarily or even partially on genetic considerations?

See Tallbear, K., and Blonick, D. A. (2004). "Native American" DNA Tests: What Are the Risks to Tribes? *The Native Voice*, Dec. 3–17.

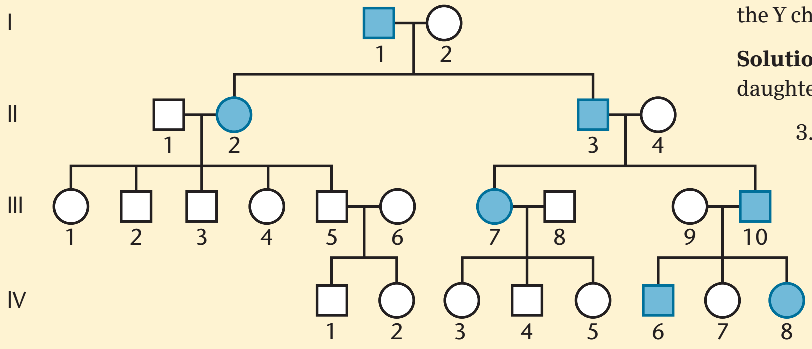
Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Patterns of inheritance sometimes vary from those expected from the biparental transmission of nuclear genes. Often, phenotypes appear to result from genetic information transmitted through the ooplasm of the egg.
2. Organelle heredity is based on the genotypes of chloroplast and mitochondrial DNA, as these organelles are transmitted to offspring. Chloroplast mutations affect the photosynthetic capabilities of plants, whereas mitochondrial mutations affect cells highly dependent on energy (ATP) generated through cellular respiration. The resulting mutants display phenotypes related to the loss of function of these organelles.
3. Both chloroplasts and mitochondria first appeared in primitive eukaryotic cells some 2 billion years ago, originating as invading protobacteria, which then coevolved with the host cell according to the endosymbiotic theory. Evidence in support of this endosymbiotic theory is extensive and centers around many observations involving the DNA and genetic machinery present in modern-day chloroplasts and mitochondria.
4. Mutations in human mtDNA are the underlying causes of a range of heritable genetic disorders in humans.
5. New therapeutic approaches are being developed to prevent the transmission of mtDNA mutations from an affected mother to her offspring.
6. Maternal-effect patterns result when nuclear gene products expressed by the maternal genotype of the egg influence early development. Coiling in snails and gene expression during early development in *Drosophila* are examples.

INSIGHTS AND SOLUTIONS

1. Analyze the following hypothetical pedigree, determine the most consistent interpretation of how the trait is inherited, and point out any inconsistencies:



Solution: The trait is passed from all affected male parents to all but one offspring, but it is *never* passed maternally. Individual IV-7 (a female) is the only exception.

2. Can the explanation in Solution 1 be attributed to a gene on the Y chromosome? Defend your answer.

Solution: No, because male parents pass the trait to their daughters as well as to their sons.

3. Is the above pedigree an example of a paternal effect or of paternal inheritance?

Solution: It has all the characteristics of paternal inheritance because males pass the trait to almost all of their offspring. To assess whether the trait is due to a paternal effect (resulting from a nuclear gene in the male gamete), analysis of further matings would be needed.

Problems and Discussion Questions

1. HOW DO WE KNOW? In this chapter, we focused on extranuclear inheritance and how traits can be determined by genetic information contained in mitochondria and chloroplasts, and we discussed how expression of maternal genotypes can affect the phenotype of an organism. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) How was it established that particular phenotypes are inherited as a result of genetic information present in the chloroplast rather than in the nucleus?
(b) How did the discovery of three categories of *petite* mutations in yeast lead researchers to postulate extranuclear inheritance of colony size?
(c) What observations support the endosymbiotic theory?
(d) What key observations in crosses between dextrally and sinistrally coiled snails support the explanation that this phenotype is the result of maternal-effect inheritance?
(e) What findings demonstrate a maternal effect as the basis of a mode of inheritance?

2. CONCEPT QUESTION Review the Chapter Concepts list on page 196. The first three center around extranuclear inheritance involving DNA in organelles. The fourth involves what is called maternal effect. Write a short essay that distinguishes between organelle heredity and maternal effect.

3. Streptomycin resistance in *Chlamydomonas* may result from a mutation in either a chloroplast gene or a nuclear gene. What phenotypic results would occur in a cross between a member of an mt^+ strain resistant in both genes and a member of a strain sensitive to the antibiotic? What results would occur in the reciprocal cross?

4. A plant may have green, white, or green-and-white (variegated) leaves on its branches, owing to a mutation in the chloroplast that prevents color from developing. Predict the results of the following crosses:

Ovule Source		Pollen Source
(a) Green branch	×	White branch
(b) White branch	×	Green branch
(c) Variegated branch	×	Green branch
(d) Green branch	×	Variegated branch

5. In diploid yeast strains, sporulation and subsequent meiosis can produce haploid ascospores, which may fuse to reestablish diploid cells. When ascospores from a *segregational petite* strain fuse with those of a normal wild-type strain, the diploid zygotes are all normal. Following meiosis, ascospores are *petite* and normal. Is the *segregational petite* phenotype inherited as a dominant or a recessive trait?

6. Predict the results of a cross between ascospores from a *segregational petite* strain and a *neutral petite* strain. Indicate the phenotype of the zygote and the ascospores it may subsequently produce.

7. In *Lymnaea*, what results would you expect in a cross between a *Dd* dextrally coiled and a *Dd* sinistrally coiled snail, assuming cross-fertilization occurs as shown in Figure 9.11? What results would occur if the *Dd* dextral produced only eggs and the *Dd* sinistral produced only sperm?

8. In a cross of *Lymnaea*, the snail contributing the eggs was dextral but of unknown genotype. Both the genotype and the phenotype of the other snail are unknown. All F_1 offspring exhibited dextral coiling. Ten of the F_1 snails were allowed to undergo self-fertilization. One-half produced only dextrally coiled offspring, whereas the other half produced only sinistrally coiled offspring. What were the genotypes of the original parents?

9. In *Drosophila subobscura*, the presence of a recessive gene called *grandchildless* (*gs*) causes the offspring of homozygous females, but not those of homozygous males, to be sterile. Can you offer an explanation as to why females and not males are affected by the mutant gene?

10. A male mouse from a true-breeding strain of hyperactive animals is crossed with a female mouse from a true-breeding strain of lethargic animals. (These are both hypothetical strains.) All the progeny are lethargic. In the F_2 generation, all offspring are lethargic. What is the best genetic explanation for these observations? Propose a cross to test your explanation.

11. Consider the case where a mutation occurs that disrupts translation in a single human mitochondrion found in the oocyte participating in fertilization. What is the likely impact of this mutation on the offspring arising from this oocyte?

12. What is the endosymbiotic theory, and why is this theory relevant to the study of extranuclear DNA in eukaryotic organelles?

13. In an earlier Problems and Discussion section (see Chapter 7, Problem 27), we described CC, the cat created by nuclear transfer cloning, whereby a diploid nucleus from one cell is injected into an enucleated egg cell to create an embryo. Cattle, sheep, rats, dogs, and several other species have been cloned using nuclei from somatic cells. Embryos and adults produced by this approach often show a number of different mitochondrial defects. Explain possible reasons for the prevalence of mitochondrial defects in embryos created by nuclear transfer cloning.

14. Mitochondrial replacement therapy (MRT) offers a potential solution for women with mtDNA-based diseases to have healthy children. Based on what you know about the importance of nuclear gene products to mitochondrial functions, will MRT ensure that children will not inherit or develop a mtDNA-based diseases?

Extra-Spicy Problems

15. The specification of the anterior–posterior axis in *Drosophila* embryos is initially controlled by various gene products that are synthesized and stored in the mature egg following oogenesis. Mutations in these genes result in abnormalities of the axis during embryogenesis. These mutations illustrate *maternal effect*. How do such mutations vary from those produced by organelle heredity? Devise a set of parallel crosses and expected outcomes involving mutant genes that contrast maternal effect and organelle heredity.

16. The maternal-effect mutation *bicoid* (*bcd*) is recessive. In the absence of the bicoid protein product, embryogenesis is not completed. Consider a cross between a female heterozygous for the *bicoid* alleles (bcd^+/bcd^-) and a male homozygous for the mutation (bcd^-/bcd^-).

(a) How is it possible for a male homozygous for the mutation to exist?

(b) Predict the outcome (normal vs. failed embryogenesis) in the F_1 and F_2 generations of the cross described.

17. (a) In humans the mitochondrial genome encodes a low number of proteins, rRNAs, and tRNAs but imports approximately 1100 proteins encoded by the nuclear genome. Yet, with such a small proportion from the mitochondrial genome encoding proteins and RNAs, a disproportionately high number of genetic disorders due to mtDNA mutations have been identified [Bigger, B. et al. (1999)]. What inheritance pattern would you expect in a three-generation pedigree in which the grandfather expresses the initial mtDNA defect? What inheritance pattern would you expect in a three-generation pedigree in which the grandmother expresses the initial mtDNA defect? (b) Considering the description in part (a) above, how would your pedigrees change if you knew that the mutation that caused the mitochondrial defect was recessive and located in the nuclear genome, was successfully transported into mitochondria, and negated a physiologically important mitochondrial function?

18. Mutations in mitochondrial DNA appear to be responsible for a number of neurological disorders, including myoclonic epilepsy and ragged-red fiber disease, Leber's hereditary optic neuropathy, and Kearns-Sayre syndrome. In each case, the disease phenotype is expressed when the ratio of mutant to wild-type mitochondria exceeds a threshold peculiar to each disease, but usually in the 60 to 95 percent range.

(a) Given that these are debilitating conditions, why has no cure been developed? Can you suggest a general approach that might be used to treat, or perhaps even cure, these disorders?

(b) Compared with the vast number of mitochondria in an embryo, the number of mitochondria in an ovum is relatively small. Might such an ooplasmic mitochondrial bottleneck present an opportunity for therapy or cure? Explain.

19. Researchers examined a family with an interesting distribution of Leigh syndrome symptoms. In this disorder, individuals may show a progressive loss of motor function (ataxia, A) with peripheral neuropathy (PN, meaning impairment of the peripheral nerves). A mitochondrial DNA (mtDNA) mutation that reduces ATPase activity was identified in various tissues of affected individuals. The accompanying table summarizes the presence of symptoms in an extended family.

Person	Condition	Percent Mitochondria with Mutation
Proband	A and PN	>90%
Brother	A and PN	>90%
Brother	Asymptomatic	17%
Mother	PN	86%
Maternal uncle	PN	85%
Maternal cousin	A and PN	90%
Maternal cousin	A and PN	91%
Maternal grandmother	Asymptomatic	56%

(a) Develop a pedigree that summarizes the information presented in the table.

(b) Provide an explanation for the pattern of inheritance of the disease. What term describes this pattern?

(c) How can some individuals in the same family show such variation in symptoms? What term, as related to organelle heredity, describes such variation?

(d) In what way does a condition caused by mtDNA differ in expression and transmission from a mutation that causes albinism?

20. Payne, B. A. et al. (2013) present evidence that a low level of heteroplasmic mtDNA exists in all tested healthy individuals.

(a) What are two likely sources of such heteroplasmy?

(b) What genetic conditions within a given mitochondrion are likely to contribute to such a variable pool of mitochondria?

21. As mentioned in Section 9.3, mtDNA accumulates mutations at a rate approximately ten times faster than nuclear DNA. Thus geneticists can use mtDNA variations as a "molecular clock" to study genetic variation and the movement of ancestral human populations from Africa to different areas of the world more than 125,000 years ago. Propose an explanation for how an analysis of mtDNA can be used to construct family trees of human evolution.

22. Because offspring inherit the mitochondrial genome only from the mother, evolutionarily the mitochondrial genome in males encounters a dead end. The mitochondrial genome in males has no significant impact on the genetic information of future generations. Scientists have proposed that this can result in an accumulation of mutations that have a negative impact on genetic fitness of males but not females. Experiments with *Drosophila* support this possibility. What experimental data or evidence would you want to evaluate or consider to determine if an accumulation of mtDNA mutations negatively impacts the fitness of males of any species?

10

DNA Structure and Analysis

Bronze sculpture of the Watson–Crick model for double-helical DNA.

CHAPTER CONCEPTS

- Except in some viruses, DNA serves as the genetic material in all living organisms on Earth.
- According to the Watson–Crick model, DNA exists in the form of a right-handed double helix.
- The strands of the double helix are antiparallel and are held together by hydrogen bonding between complementary nitrogenous bases.
- The structure of DNA provides the means of storing and expressing genetic information.
- RNA has many similarities to DNA but exists mostly as a single-stranded molecule.
- In some viruses, RNA serves as the genetic material.
- Many techniques have been developed that facilitate the analysis of nucleic acids, most based on detection of the complementarity of nitrogenous bases.

Up to this point in the text, we have described chromosomes as containing genes that control phenotypic traits transmitted through gametes to future offspring. Logically, genes must contain some sort of information that, when passed to a new generation, influences the form and characteristics of each individual. We refer to that information as the **genetic material**. Logic also suggests that this same information in some way directs the many complex processes that lead to an organism's adult form.

Until 1944, it was not clear what chemical component of the chromosome makes up genes and constitutes the genetic material. Because chromosomes were known to have both a nucleic acid and a protein component, both were candidates. In 1944, however, direct experimental evidence emerged showing that the nucleic acid DNA serves as the informational basis for the process of heredity.

Once the importance of DNA to genetic processes was realized, work was intensified with the hope of discerning not only the structure of this molecule but also the relationship of its structure to its function. Between 1944 and 1953, many scientists sought information that might answer the most significant and intriguing question in the history of biology: How does DNA serve as the genetic basis for living processes? Researchers believed the answer must depend strongly on the chemical structure of the DNA molecule, given the complex but orderly functions ascribed to it.

These efforts were rewarded in 1953, when James Watson and Francis Crick put forth their hypothesis for the double-helical nature of DNA. The assumption that the molecule's functions would be easier to clarify once its general structure was determined proved to be correct. In this chapter, we first review the evidence that DNA is the genetic material and then discuss the elucidation of its structure. We conclude the chapter with a discussion of various analytical techniques useful during the investigation of the nucleic acids, DNA and RNA.

10.1 The Genetic Material Must Exhibit Four Characteristics

For a molecule to serve as the genetic material, it must exhibit four crucial characteristics: **replication, storage of information, expression of information,** and **variation by mutation.** *Replication* of the genetic material is one facet of the cell cycle and as such is a fundamental property of all living organisms. Once the genetic material of cells replicates and is doubled in amount, it must then be partitioned equally—through mitosis—into daughter cells. The genetic material is also replicated during the formation of gametes, but is partitioned so that each cell gets only one-half of the original amount of genetic material—the process of *meiosis* (discussed in Chapter 2). Although the products of mitosis and meiosis are different, these processes are both part of the more general phenomenon of cellular reproduction.

Storage of information requires the molecule to act as a repository of genetic information that may or may not be expressed by the cell in which it resides. It is clear that while most cells contain a complete copy of the organism's genome, at any point in time they express only a part of this genetic potential. For example, in bacteria many genes "turn on" in response to specific environmental cues and "turn off" when conditions change. In vertebrates, skin cells may display active melanin genes but never activate their hemoglobin genes; in contrast, digestive cells activate many genes specific to their function but do not activate their melanin genes.

Inherent in the concept of storage is the need for the genetic material to be able to encode the vast variety of gene products found among the countless forms of life on our planet. The chemical language of the genetic material must have the capability of storing such diverse information and transmitting it to progeny cells and organisms.

Expression of the stored genetic information is a complex process that is the underlying basis for the concept of **information flow** within the cell (**Figure 10.1**). The initial event in this flow of information is the **transcription** of DNA, in which three main types of RNA molecules are synthesized: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Of these, mRNAs are translated into proteins, by means of a process mediated by the tRNA and rRNA. Each mRNA is the product of a specific gene and leads to the synthesis of a different protein. In **translation,** the chemical information in mRNA directs the construction of a chain of amino acids, called a polypeptide, which then folds into a protein. Collectively, these processes serve as the foundation for the **central dogma of molecular genetics**: "DNA makes RNA, which makes proteins."

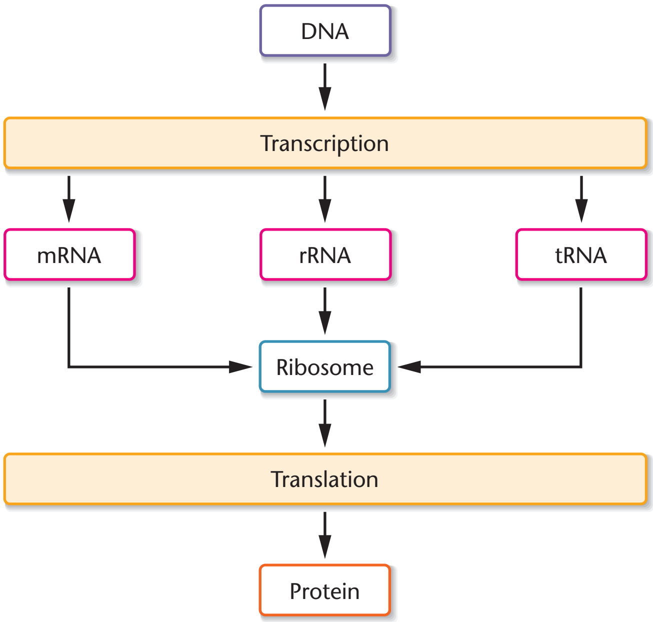


FIGURE 10.1 Simplified diagram of information flow (the central dogma) from DNA to RNA to produce the proteins within cells.

The genetic material is also the source of *variability* among organisms, through the process of mutation. If a mutation—a change in the chemical composition of DNA—occurs, the alteration is reflected during transcription and translation, affecting the specific protein. If a mutation is present in a gamete, it may be passed to future generations and, with time, become distributed in the population. Genetic variation, which also includes alterations of chromosome number and rearrangements within and between chromosomes (as discussed in Chapter 8), provides the raw material for the process of evolution.

10.2 Until 1944, Observations Favored Protein as the Genetic Material

The idea that genetic material is physically transmitted from parent to offspring has been accepted for as long as the concept of inheritance has existed. Beginning in the late nineteenth century, research into the structure of biomolecules progressed considerably, setting the stage for describing the genetic material in chemical terms. Although proteins and nucleic acid were both considered major candidates for the role of genetic material, until the 1940s many geneticists favored proteins. This is not surprising, since proteins were known to be both diverse and abundant in cells, and much more was known about protein than about nucleic acid chemistry.

DNA was first studied in 1869 by a Swiss chemist, Friedrich Miescher. He isolated cell nuclei and derived an acidic substance, now known to contain DNA, that he called **nuclein**. As investigations of DNA progressed, however, showing it to be present in chromosomes, the substance seemed to lack the chemical diversity necessary to store extensive genetic information.

This conclusion was based largely on Phoebus A. Levene's observations in 1910 that DNA contained approximately equal amounts of four similar molecules called *nucleotides*. Levene postulated incorrectly that identical groups of these four components were repeated over and over, which was the basis of his **tetranucleotide hypothesis** for DNA structure. Attention was thus directed away from DNA, thereby favoring proteins. However, in the 1940s, Erwin Chargaff showed that Levene's proposal was incorrect when he demonstrated that most organisms do not contain precisely equal proportions of the four nucleotides. We shall see later that the structure of DNA accounts for Chargaff's observations.

10.3 Evidence Favoring DNA as the Genetic Material Was First Obtained during the Study of Bacteria and Bacteriophages

The 1944 publication by Oswald Avery, Colin MacLeod, and Maclyn McCarty concerning the chemical nature of a "transforming principle" in bacteria was the initial event leading to the acceptance of DNA as the genetic material. Their work, along with subsequent findings of other research teams, constituted the first direct experimental proof that DNA, and not protein, is the biomolecule responsible for heredity. It marked the beginning of the *era of molecular genetics*, a period of discovery in biology that made biotechnology feasible and has moved us closer to an understanding of the basis of life. The impact of their initial findings on future research and thinking paralleled that of the publication of Darwin's theory of evolution and the subsequent rediscovery of Mendel's postulates of transmission genetics. Together, these events constitute three great revolutions in biology.

Transformation: Early Studies

The research that provided the foundation for Avery, MacLeod, and McCarty's work was initiated in 1927 by Frederick Griffith, a medical officer in the British Ministry of Health. He performed experiments with several different strains of the bacterium *Diplococcus pneumoniae*.* Some were *virulent*, that is, infectious, strains that cause pneumonia in certain vertebrates (notably humans and mice), whereas others were *avirulent*, or noninfectious strains, which do not cause illness.

The difference in virulence depends on the presence of a polysaccharide capsule; virulent strains have this capsule, whereas avirulent strains do not. The nonencapsulated bacteria are readily engulfed and destroyed by phagocytic cells in the host animal's circulatory system. Virulent bacteria, which possess the polysaccharide coat, are not easily engulfed; they multiply and cause pneumonia.

The presence or absence of the capsule causes a visible difference between colonies of virulent and avirulent strains. Encapsulated bacteria form smooth, shiny-surfaced colonies (*S*) when grown on an agar culture plate; nonencapsulated strains produce rough colonies (*R*). Thus, virulent and avirulent strains are easily distinguished by standard microbiological culture techniques.

Each strain of *Diplococcus* may be one of dozens of different types called *serotypes* that differ in the precise chemical structure of the polysaccharide constituent of the thick, slimy capsule. Serotypes are identified by immunological techniques and are usually designated by Roman numerals. In the United States, types I and II are the most common in causing pneumonia. Griffith used types II*R* and III*S* in his critical experiments that led to new concepts about the genetic material. **Table 10.1** summarizes the characteristics of Griffith's two strains, while **Figure 10.2** on p. 216 depicts his experiment.

Griffith knew from the work of others that only living virulent cells would produce pneumonia in mice. If heat-killed virulent bacteria are injected into mice, no pneumonia results, just as living avirulent bacteria fail to produce the disease. Griffith's critical experiment (Figure 10.2) involved an injection into mice of living II*R* (avirulent) cells combined with heat-killed III*S* (virulent) cells. Since neither cell type caused death in mice when injected alone, Griffith expected that the double injection would not kill the mice. But, after five days, all of the mice that received both types of cells were dead. Paradoxically, analysis of their blood revealed a large number of living type III*S* (virulent) bacteria.

As far as could be determined, these III*S* bacteria were identical to the III*S* strain from which the heat-killed cell

TABLE 10.1 Strains of *Diplococcus pneumonia* Used by Frederick Griffith in His Original Transformation Experiments

Serotype	Colony Morphology	Capsule	Virulence
II*R*	Rough	Absent	Avirulent
III*S*	Smooth	Present	Virulent

*This organism is now named *Streptococcus pneumoniae*.

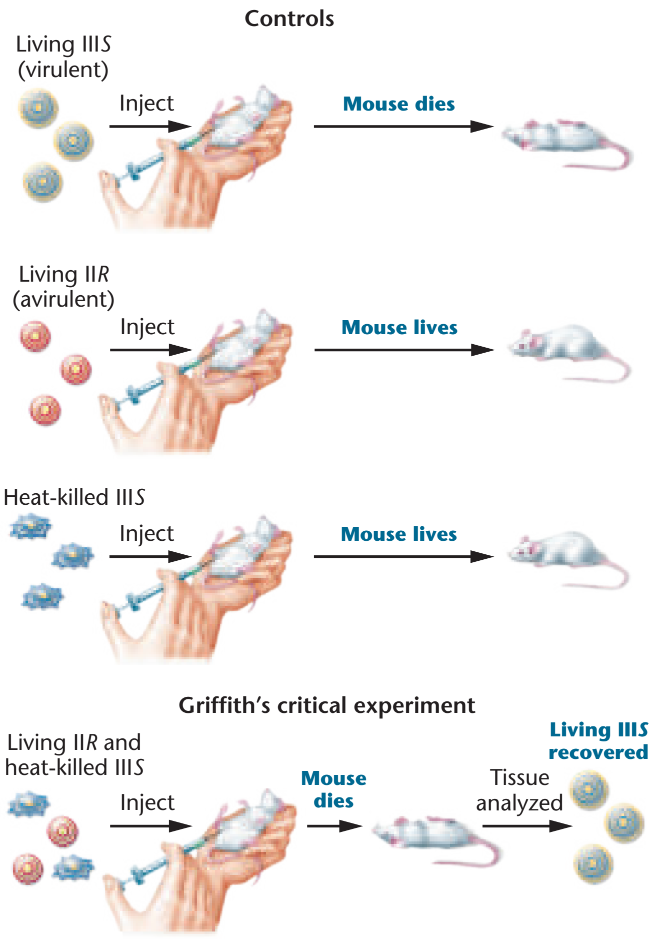


FIGURE 10.2 Griffith's transformation experiment.

preparation had been made. The control mice, injected only with living avirulent II*R* bacteria for this set of experiments, did not develop pneumonia and remained healthy. This ruled out the possibility that the avirulent II*R* cells simply changed (or mutated) to virulent III*S* cells in the absence of the heat-killed III*S* bacteria. Instead, some type of interaction had taken place between living II*R* and heat-killed III*S* cells.

Griffith concluded that the heat-killed III*S* bacteria somehow converted live avirulent II*R* cells into virulent III*S* cells. Calling the phenomenon **transformation,** he suggested that the *transforming principle* might be some part of the polysaccharide capsule or a compound required for capsule synthesis, although the capsule alone did not cause pneumonia. To use Griffith's term, the transforming principle from the dead III*S* cells served as a "pabulum"— that is, a nutrient source—for the II*R* cells.

Griffith's work led other physicians and bacteriologists to research the phenomenon of transformation. By 1931, M. Henry Dawson at the Rockefeller Institute had confirmed Griffith's observations and extended his work one step further. Dawson and his coworkers showed that transformation could occur *in vitro* (in a test tube). When heat-killed III*S* cells were incubated with living II*R* cells, living III*S* cells were recovered. Therefore, injection into mice was not necessary for transformation to occur. By 1933, J. Lionel Alloway had refined the *in vitro* experiments by using crude extracts of III*S* cells and living II*R* cells. The soluble filtrate from the heat-killed III*S* cells was as effective in inducing transformation as were the intact cells. Alloway and others did not view transformation as a genetic event, but rather as a physiological modification of some sort. Nevertheless, the experimental evidence that a chemical substance was responsible for transformation was quite convincing.

Transformation: The Avery, MacLeod, and McCarty Experiment

The critical question, of course, was what molecule serves as the transforming principle? In 1944, after 10 years of work, Avery, MacLeod, and McCarty published their results in what is now regarded as a classic paper in the field of molecular genetics. They reported that they had obtained the transforming principle in a purified state and that beyond reasonable doubt it was DNA.

The details of their work, sometimes called the Avery, MacLeod, and McCarty experiment, are outlined in Figure 10.3. These researchers began their isolation procedure with large quantities (50–75 liters) of liquid cultures of type III*S* virulent cells. The cells were centrifuged, collected, and heat killed. Following homogenization and several extractions with the detergent deoxycholate (DOC), the researchers obtained a soluble filtrate that retained the ability to induce transformation of type II*R* avirulent cells. Protein was removed from the active filtrate by several chloroform extractions, and polysaccharides were enzymatically digested and removed. Finally, precipitation with ethanol yielded a fibrous mass that still retained the ability to induce transformation of type II*R* avirulent cells. From the original 75-liter sample, the procedure yielded 10–25 mg of this "active factor."

Further testing clearly established that the transforming principle was DNA. The fibrous mass was first analyzed for its nitrogen:phosphorus ratio, which was shown to coincide with the ratio of "sodium desoxyribonucleate," the chemical name then used to describe DNA. To solidify their findings, Avery, MacLeod, and McCarty sought to eliminate, to the greatest extent possible, all probable contaminants from their final product. Thus, it was treated with the proteolytic enzymes trypsin and chymotrypsin and then with an RNA-digesting enzyme, called **ribonuclease (RNase).** Such treatments destroyed any remaining activity of proteins and RNA. Nevertheless, transforming activity still remained. Chemical testing of the final product gave strong positive reactions for DNA. The final confirmation came with experiments using crude samples of the DNA-digesting enzyme **deoxyribonuclease (DNase),** which

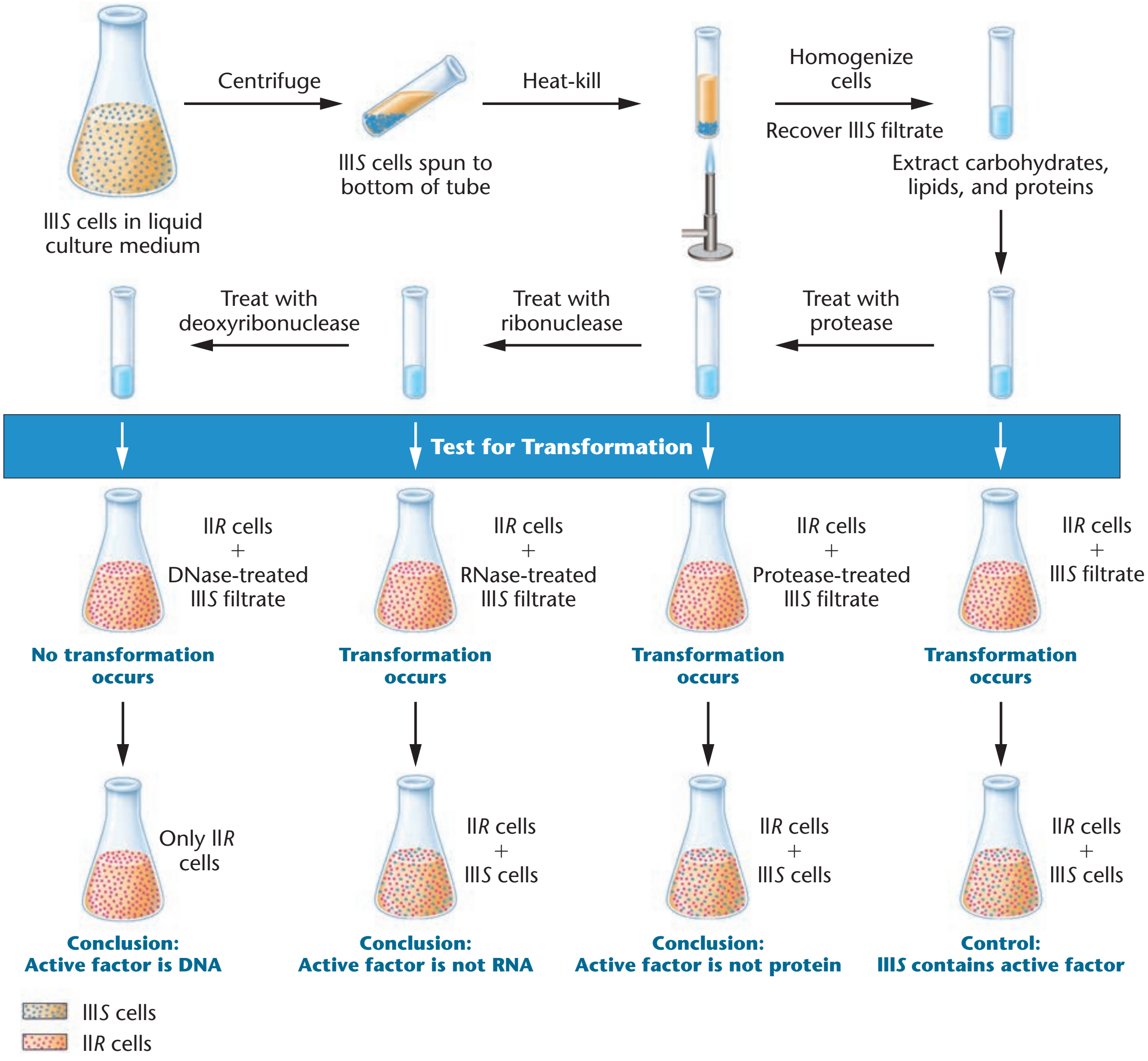

Centrifuge
Heat-kill
Homogenize cells
Recover IIIS filtrate
IIIS cells in liquid culture medium
IIIS cells spun to bottom of tube
Extract carbohydrates, lipids, and proteins
Treat with deoxyribonuclease
Treat with ribonuclease
Treat with protease
Test for Transformation
IIR cells + DNase-treated IIIS filtrate
IIR cells + RNase-treated IIIS filtrate
IIR cells + Protease-treated IIIS filtrate
IIR cells + IIIS filtrate
No transformation occurs
Transformation occurs
Transformation occurs
Transformation occurs
Only IIR cells
IIR cells + IIIS cells
IIR cells + IIIS cells
IIR cells + IIIS cells
Conclusion: Active factor is DNA
Conclusion: Active factor is not RNA
Conclusion: Active factor is not protein
Control: IIIS contains active factor
IIIS cells
IIR cells

FIGURE 10.3 Summary of Avery, MacLeod, and McCarty's experiment demonstrating that DNA is the transforming principle.

was isolated from dog and rabbit sera. Digestion with this enzyme destroyed the transforming activity of the filtrate—thus Avery and his coworkers were certain that the active transforming principle in these experiments was DNA.

The great amount of work involved in this research, the confirmation and reconfirmation of the conclusions drawn, and the unambiguous logic of the experimental design are truly impressive. Avery, MacLeod, and McCarty's conclusion in the 1944 publication was, however, very simply stated: "The evidence presented supports the belief that a nucleic acid of the desoxyribose* type is the fundamental unit of the transforming principle of *Pneumococcus* Type III."

Avery and his colleagues recognized the genetic and biochemical implications of their work. They observed that "nucleic acids of this type must be regarded not merely as structurally important but as functionally active in determining the biochemical activities and specific characteristics of pneumococcal cells." This suggested that the transforming principle interacts with the II*R* cell and gives rise to a coordinated series of enzymatic reactions culminating in the synthesis of the type III*S* capsular polysaccharide. Avery, MacLeod, and McCarty emphasized that, once transformation occurs, the capsular polysaccharide is produced in successive generations. Transformation is therefore heritable, and the process affects the genetic material.

Transformation has now been shown to occur in *Haemophilus influenzae, Bacillus subtilis, Shigella paradysenteriae*, and *Escherichia coli*, among many other microorganisms. Transformation of numerous genetic traits other than colony morphology has also been demonstrated, including traits involving resistance to antibiotics. These observations further strengthened the belief that transformation by DNA

*Desoxyribose is now spelled deoxyribose.

is primarily a genetic event rather than simply a physiological change. We will pursue this idea again in the "Insights and Solutions" section at the end of this chapter.

The Hershey–Chase Experiment

The second major piece of evidence supporting DNA as the genetic material was provided during the study of the bacterium *Escherichia coli* and one of its infecting viruses, **bacteriophage T2.** Often referred to simply as a **phage,** the virus consists of a protein coat surrounding a core of DNA. Electron micrographs reveal that the phage's external structure is composed of a icosahedral head plus a tail. **Figure 10.4** shows as much of the life cycle as was known in 1952 for a T-even bacteriophage such as T2. Briefly, the phage adsorbs to the bacterial cell, and some genetic component of the phage enters the bacterial cell. Following infection, the viral component "commandeers" the cellular machinery of the host and causes viral reproduction. In a reasonably short time, many new phages are constructed and the bacterial cell is lysed, releasing the progeny viruses. This process is referred to as the **lytic cycle**.

In 1952, Alfred Hershey and Martha Chase published the results of experiments designed to clarify the events leading to phage reproduction. Several of the experiments clearly established the independent functions of phage protein and nucleic acid in the reproduction process associated with the bacterial cell. Hershey and Chase knew from existing data that:

1. T2 phages consist of approximately 50 percent protein and 50 percent DNA.
2. Infection is initiated by adsorption of the phage by its tail fibers to the bacterial cell.
3. The production of new viruses occurs within the bacterial cell.

It appeared that some molecular component of the phage—DNA or protein (or both)—entered the bacterial cell and directed viral reproduction. Which was it?

Hershey and Chase used the radioisotopes ^{32}P and ^{35}S to follow the molecular components of phages during infection. Because DNA contains phosphorus (P) but not sulfur, ^{32}P effectively labels DNA; because proteins contain sulfur (S) but not phosphorus, ^{35}S labels protein. *This is a key feature of the experiment*. If *E. coli* cells are first grown in the presence of ^{32}P *or* ^{35}S and then infected with T2 viruses, the progeny phages will have *either* a radioactively labeled

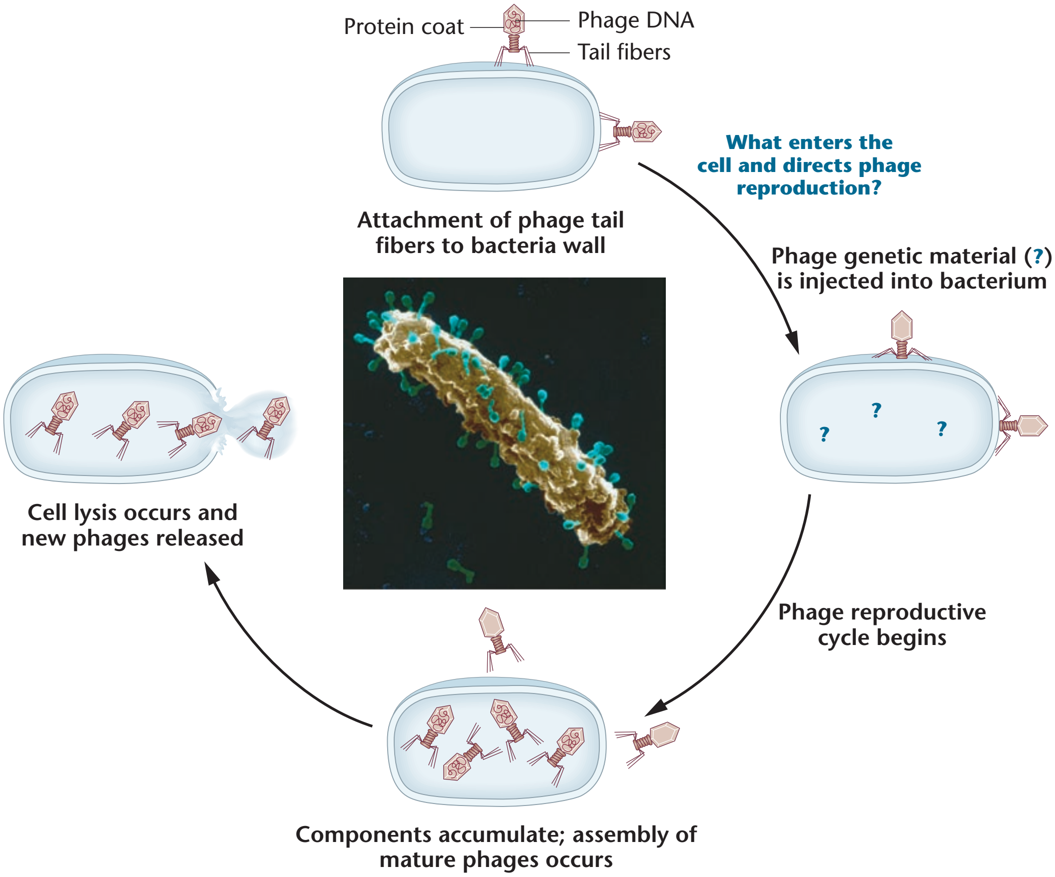

Protein coat
Phage DNA
Tail fibers
Attachment of phage tail fibers to bacteria wall
What enters the cell and directs phage reproduction?
Phage genetic material (?) is injected into bacterium
?
?
?
Phage reproductive cycle begins
Components accumulate; assembly of mature phages occurs
Cell lysis occurs and new phages released

FIGURE 10.4 Life cycle of a T-even bacteriophage, as known in 1952. The electron micrograph shows an *E. coli* cell during infection by numerous T2 phages (shown in blue).

DNA core *or* a radioactively labeled protein coat, respectively. These labeled phages can be isolated and used to infect unlabeled bacteria (**Figure 10.5**).

When labeled phages and unlabeled bacteria were mixed, an adsorption complex was formed as the phages attached their tail fibers to the bacterial wall. These complexes were isolated and subjected to a high shear force in a blender. The force stripped off the attached phages so that the phages and bacteria could be analyzed separately. Centrifugation separated the lighter phage particles from the heavier bacterial cells (Figure 10.5). By tracing the radioisotopes, Hershey and Chase were able to demonstrate that most of the ^{32}P-labeled DNA had been transferred into the bacterial cell following adsorption; on the other hand, almost all of the ^{35}S-labeled protein remained outside the bacterial cell and was recovered in the phage "ghosts" (empty phage coats) after the blender treatment. Following this separation, the bacterial cells, which now contained viral DNA, were eventually lysed as new phages were produced. These progeny phages contained ^{32}P, but not ^{35}S.

Hershey and Chase interpreted these results as indicating that the protein of the phage coat remains outside the

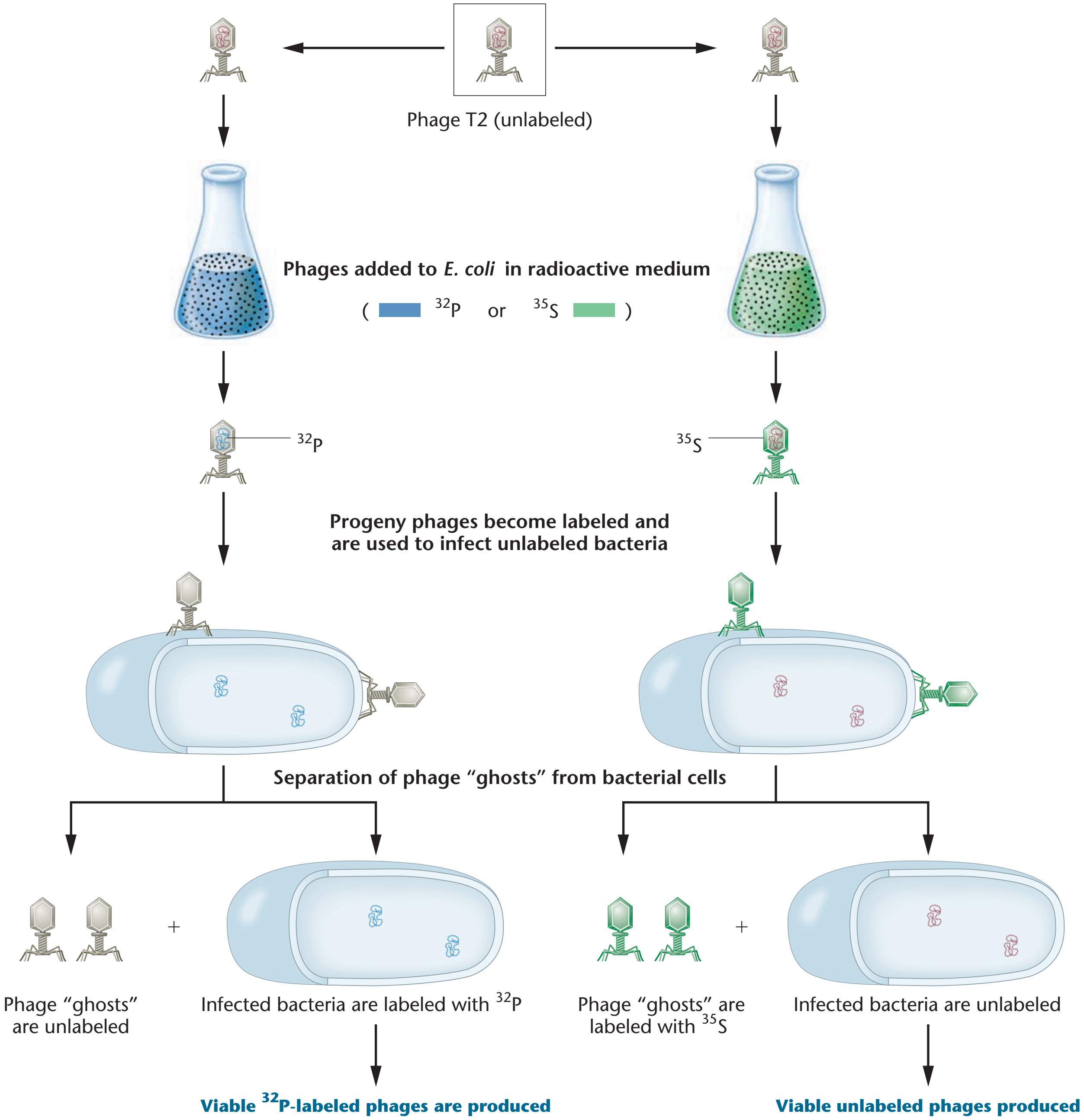

Phage T2 (unlabeled)
Phages added to *E. coli* in radioactive medium
(🟦 ^{32}P or ^{35}S 🟩)
^{32}P
^{35}S
Progeny phages become labeled and are used to infect unlabeled bacteria
Separation of phage "ghosts" from bacterial cells
+
Phage "ghosts" are unlabeled
Infected bacteria are labeled with ^{32}P
+
Phage "ghosts" are labeled with ^{35}S
Infected bacteria are unlabeled
Viable ^{32}P-labeled phages are produced
Viable unlabeled phages produced

FIGURE 10.5 Summary of the Hershey–Chase experiment demonstrating that DNA, and not protein, is responsible for directing the reproduction of phage T2 during the infection of *E. coli*.

host cell and is not involved in directing the production of new phages. On the other hand, and most important, phage DNA enters the host cell and directs phage reproduction. Hershey and Chase had demonstrated that the genetic material in phage T2 is DNA, not protein.

These experiments, along with those of Avery and his colleagues, provided convincing evidence that DNA was the molecule responsible for heredity. This conclusion has since served as the cornerstone of the field of molecular genetics.

NOW SOLVE THIS

10.1 Would an experiment similar to that performed by Hershey and Chase work if the basic design were applied to the phenomenon of transformation? Explain why or why not.

- **HINT:** *This problem involves an understanding of the protocol of the Hershey–Chase experiment as applied to the investigation of transformation. The key to its solution is to remember that in transformation, exogenous DNA enters the soon-to-be transformed cell and that no cell-to-cell contact is involved in the process.*

Transfection Experiments

During the eight years following publication of the Hershey–Chase experiment, additional research using bacterial viruses provided even more solid proof that DNA is the genetic material. In 1957, several reports demonstrated that if *E. coli* is treated with the enzyme lysozyme, the outer wall of the cell can be removed without destroying the bacterium. Enzymatically treated cells are naked, so to speak, and contain only the cell membrane as their outer boundary. Such structures are called **protoplasts** (or **spheroplasts**). John Spizizen and Dean Fraser independently reported that by using protoplasts, they were able to initiate phage reproduction with DNA derived from T2 phages. That is, provided protoplasts were used, a virus did not have to be intact for infection to occur. Thus, the outer protein coat structure may be essential to the movement of DNA through the intact cell wall, but it is not essential for infection when protoplasts are used.

Similar, but more refined, experiments were reported in 1960 by George Guthrie and Robert Sinsheimer. DNA was purified from bacteriophawwge ϕX174, a small phage that contains a single-stranded circular DNA molecule of some 5386 nucleotides. When added to *E. coli* protoplasts, the purified DNA resulted in the production of complete ϕX174 bacteriophages. This process of infection by only the viral nucleic acid, called **transfection,** proves conclusively that ϕX 174 DNA alone contains all the necessary information for production of mature viruses. Thus, the evidence that DNA serves as the genetic material was further strengthened, even though all direct evidence to that point had been obtained from bacterial and viral studies.

10.4 Indirect and Direct Evidence Supports the Concept That DNA Is the Genetic Material in Eukaryotes

In 1950, eukaryotic organisms were not amenable to the types of experiments that used bacteria and viruses to demonstrate that DNA is the genetic material. Nevertheless, it was generally assumed that the genetic material would be a universal substance serving the same role in eukaryotes. Initially, support for this assumption relied on several circumstantial observations that, taken together, indicated that DNA does serve as the genetic material in eukaryotes. Subsequently, direct evidence established unequivocally the central role of DNA in genetic processes.

Indirect Evidence: Distribution of DNA

The genetic material should be found where it functions—in the nucleus as part of chromosomes. Both DNA and protein fit this criterion. However, protein is also abundant in the cytoplasm, whereas DNA is not. Both mitochondria and chloroplasts are known to perform genetic functions, and DNA is also present in these organelles. Thus, DNA is found only where primary genetic functions occur. Protein, on the other hand, is found everywhere in the cell. These observations are consistent with the interpretation favoring DNA over proteins as the genetic material.

Because it had earlier been established that chromosomes within the nucleus contain the genetic material, a correlation was expected to exist between the ploidy (n, $2n$, etc.) of a cell and the quantity of the substance that functions as the genetic material. Meaningful comparisons can be made between gametes (sperm and eggs) and somatic or body cells. The latter are recognized as being diploid ($2n$) and containing twice the number of chromosomes as gametes, which are haploid (n).

Table 10.2 compares, for a variety of organisms, the amount of DNA found in haploid sperm to the amount found in diploid nucleated precursors of red blood cells. The amount of DNA and the number of sets of chromosomes is closely correlated. No such consistent correlation can be observed between gametes and diploid cells for proteins. These data thus provide further circumstantial evidence favoring DNA over proteins as the genetic material of eukaryotes.

TABLE 10.2 DNA Content of Haploid versus Diploid Cells of Various Species*

Organism	*n* (pg)	2*n* (pg)
Human	3.25	7.30
Chicken	1.26	2.49
Trout	2.67	5.79
Carp	1.65	3.49
Shad	0.91	1.97

*Sperm (n) and nucleated precursors to red blood cells ($2n$) were used to contrast ploidy levels.

Indirect Evidence: Mutagenesis

Ultraviolet (UV) light is one of a number of agents capable of inducing mutations in the genetic material. Simple organisms such as yeast and other fungi can be irradiated with various wavelengths of UV light and the effectiveness of each wavelength measured by the number of mutations it induces. When the data are plotted, an **action spectrum** of UV light as a mutagenic agent is obtained. This action spectrum can then be compared with the **absorption spectrum** of any molecule suspected to be the genetic material (**Figure 10.6**). *The molecule serving as the genetic material is expected to absorb at the wavelength(s) found to be mutagenic.*

UV light is most mutagenic at the wavelength of 260 nanometers (nm), and both DNA and RNA absorb UV light most strongly at 260 nm. On the other hand, protein absorbs most strongly at 280 nm, yet no significant mutagenic effects are observed at that wavelength. This indirect evidence supports the idea that a nucleic acid, rather than protein, is the genetic material.

Direct Evidence: Recombinant DNA Studies

Although the circumstantial evidence just described does not constitute direct proof that DNA is the genetic material in eukaryotes, over a half century of research has provided irrefutable evidence that DNA serves this role. In fact, this simple concept of genetics is at the foundation of modern genetic research and its applications.

For example, **recombinant DNA technology**, which involves splicing together DNA sequences from different organisms (see Chapter 20), has combined the DNA sequence encoding the human hormone insulin with bacterial DNA sequences. When this recombinant DNA molecule is introduced into bacteria, the bacteria replicate the DNA and pass it to daughter cells at each cell division. In addition, the bacteria express the recombinant DNA molecule and thus produce human insulin. This example of biotechnology clearly establishes how a specific DNA sequence confers heritable information responsible for a product of a gene.

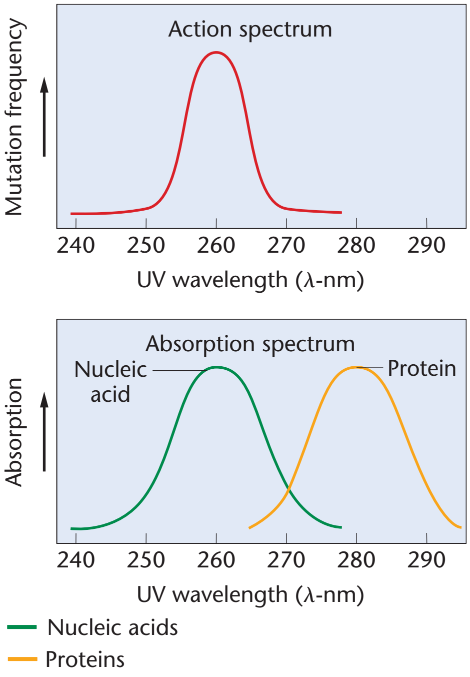


FIGURE 10.6 Comparison of the action spectrum (which determines the most effective mutagenic UV wavelength) and the absorption spectrum (which shows the range of wavelength where nucleic acids and proteins absorb UV light).

Genomics (see Chapter 21), which can provide the full set of DNA sequences of organisms, is the basis of still another example in support of this concept. We have known the full sequence of the human genome since 2001. However, geneticists are still uncovering new clues as to how the 3.2 billion base pairs of human DNA serve as the basis for human life. For example, the sequencing of genomes from individuals with specific heritable disorders and their comparison to genomes of healthy individuals has provided many insights about which DNA sequences (or genes) harbor mutations responsible for these genetic disorders. The underlying premise of such studies is that DNA is the genetic material.

By the mid-1970s the concept that DNA is the genetic material in eukaryotes was accepted, and since then, no information has been forthcoming to dispute that conclusion. In the upcoming chapters, we will see exactly how DNA is stored, replicated, mutated, repaired, and expressed.

10.5 RNA Serves as the Genetic Material in Some Viruses

Some viruses contain an RNA core rather than a DNA core. In these viruses, it appears that RNA serves as the genetic material—an exception to the general rule that DNA performs this function. In 1956, it was demonstrated that when purified RNA from **tobacco mosaic virus (TMV)** was spread on tobacco leaves, the characteristic lesions caused by viral infection subsequently appeared. Thus, it was concluded that RNA is the genetic material of this virus.

In 1965 and 1966, Norman Pace and Sol Spiegelman demonstrated that RNA from the phage Qβ can be isolated and replicated *in vitro*. Replication depends on an enzyme, **RNA replicase,** which is isolated from host *E. coli* cells following normal infection. When the RNA replicated *in vitro* is added to *E. coli* protoplasts, infection and viral multiplication (*transfection*) occur. Thus, RNA synthesized in a test tube serves as the genetic material in these phages by directing the production of all the components necessary for viral reproduction.

While many viruses, such the T2 virus used by Hershey and Chase, use DNA as their hereditary material, another group of RNA-containing viruses bears mention. These are the **retroviruses,** which replicate in an unusual way. Their RNA serves as a template for the synthesis of the complementary

DNA molecule. The process, **reverse transcription,** occurs under the direction of an RNA-dependent DNA polymerase enzyme called **reverse transcriptase.** This DNA intermediate can be incorporated into the genome of the host cell, and when the host DNA is transcribed, copies of the original retroviral RNA chromosomes are produced. Retroviruses include the human immunodeficiency virus (HIV), which causes AIDS, as well as several RNA tumor viruses.

10.6 Knowledge of Nucleic Acid Chemistry Is Essential to the Understanding of DNA Structure

Having established the critical importance of DNA and RNA in genetic processes, we will now take a brief look at the chemical structures of these molecules. As we shall see, the structural components of DNA and RNA are very similar. This chemical similarity is important in the coordinated functions played by these molecules during gene expression. Like the other major groups of organic biomolecules (proteins, carbohydrates, and lipids), nucleic acid chemistry is based on a variety of similar building blocks that are polymerized into chains of varying lengths.

Nucleotides: Building Blocks of Nucleic Acids

DNA is a nucleic acid, and **nucleotides** are the building blocks of all nucleic acid molecules. Sometimes called mononucleotides, these structural units consist of three essential components: a **nitrogenous base,** a **pentose sugar** (a 5-carbon sugar), and a **phosphate group.** There are two kinds of nitrogenous bases: the nine-member double-ring **purines** and the six-member single-ring **pyrimidines**.

Two types of purines and three types of pyrimidines are commonly found in nucleic acids. The two purines are **guanine** and **adenine,** abbreviated **G** and **A**. The three pyrimidines are **cytosine** , **thymine,** and **uracil,** abbreviated **C, T,** and **U,** respectively. The chemical structures of A, G, C, T, and U are shown in **Figure 10.7(a)**. Both DNA and

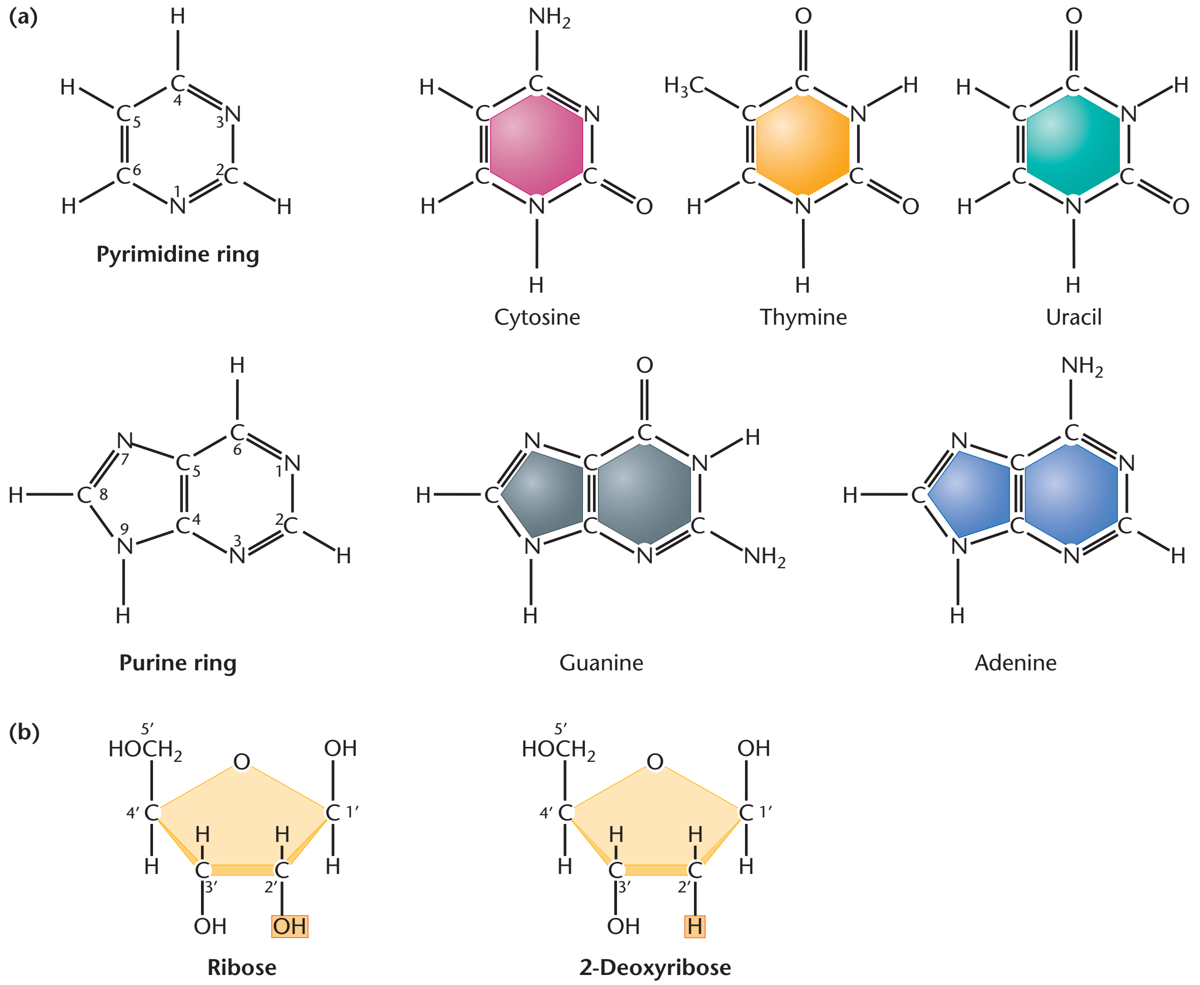


FIGURE 10.7 (a) Chemical structures of the pyrimidines and purines that serve as the nitrogenous bases in RNA and DNA. The convention for numbering carbon and nitrogen atoms making up the two categories of bases is shown within the structures that appear on the left. (b) Chemical ring structures of ribose and 2-deoxyribose, which serve as the pentose sugars in RNA and DNA, respectively.

RNA contain A, C, and G, but only DNA contains the base T and only RNA contains the base U. Each nitrogen or carbon atom of the ring structures of purines and pyrimidines is designated by an unprimed number. Note that corresponding atoms in the two rings are numbered differently in most cases.

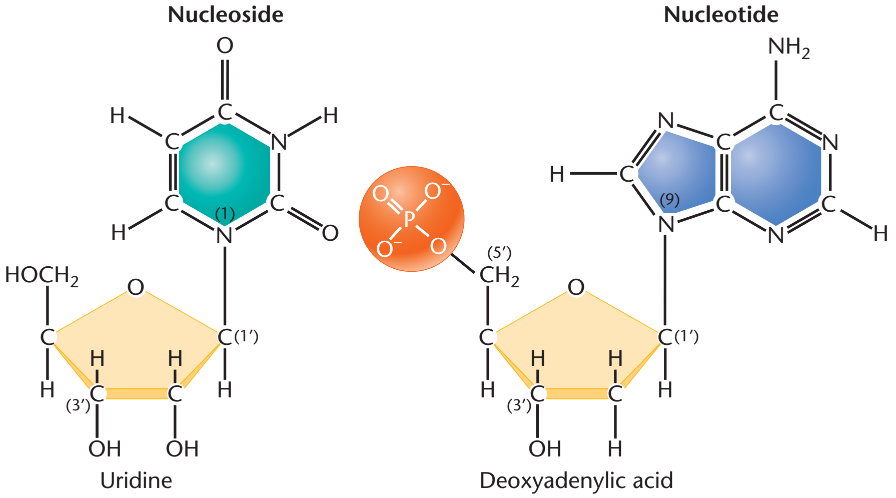
The pentose sugars found in nucleic acids give them their names. Ribonucleic acids (RNA) contain **ribose,** while deoxyribonucleic acids (DNA) contain **deoxyribose**. Figure 10.7(b) shows the ring structures for these two pentose sugars. Each carbon atom is distinguished by a number with a prime sign (e.g., C-1′, C-2′). Compared with ribose, deoxyribose has a hydrogen atom rather than a hydroxyl group at the C-2′ position. The absence of a hydroxyl group at the C-2′ position thus distinguishes DNA from RNA. In the absence of the C-2′ hydroxyl group, the sugar is more specifically named **2-deoxyribose**.

If a molecule is composed of a purine or pyrimidine base and a ribose or deoxyribose sugar, the chemical unit is called a **nucleoside**. If a phosphate group is added to the nucleoside, the molecule is now called a nucleotide. Nucleosides and nucleotides are named according to the specific nitrogenous base (A, T, G, C, or U) that is part of the molecule. The structures of a nucleoside and a nucleotide and the nomenclature used in naming nucleosides and nucleotides are given in Figure 10.8.

The bonding between components of a nucleotide is highly specific. The C-1′ atom of the sugar is involved in the chemical linkage to the nitrogenous base. If the base is a purine, the N-9 atom is covalently bonded to the sugar; if the base is a pyrimidine, the N-1 atom bonds to the sugar. In deoxyribonucleotides, the phosphate group may be bonded to the C-2′, C-3′, or C-5′ atom of the sugar. The C-5′ phosphate configuration is shown in Figure 10.8. It is by far the prevalent form in biological systems and the one found in DNA and RNA.

Nucleoside Diphosphates and Triphosphates

Nucleotides are also described by the term **nucleoside monophosphate (NMP)**. The addition of one or two phosphate groups results in **nucleoside diphosphates (NDPs)** and **triphosphates (NTPs)**, respectively, as shown in Figure 10.9. The triphosphate form is significant because it serves as the precursor molecule during nucleic acid synthesis within the cell (see Chapter 11). In addition, **adenosine triphosphate (ATP)** and **guanosine triphosphate (GTP)** are important in cell bioenergetics because of the large amount of energy involved in adding or removing the terminal phosphate group. The hydrolysis of ATP or GTP to ADP or GDP and inorganic phosphate (P_i) is accompanied by the release of a large amount of energy in the cell. When these chemical conversions are coupled to other reactions, the



Ribonucleosides	Ribonucleotides
Adenosine	Adenylic acid
Cytidine	Cytidylic acid
Guanosine	Guanylic acid
Uridine	Uridylic acid
Deoxyribonucleosides	**Deoxyribonucleotides**
Deoxyadenosine	Deoxyadenylic acid
Deoxycytidine	Deoxycytidylic acid
Deoxyguanosine	Deoxyguanylic acid
Deoxythymidine	Deoxythymidylic acid

FIGURE 10.8 Structures and names of the nucleosides and nucleotides of RNA and DNA.

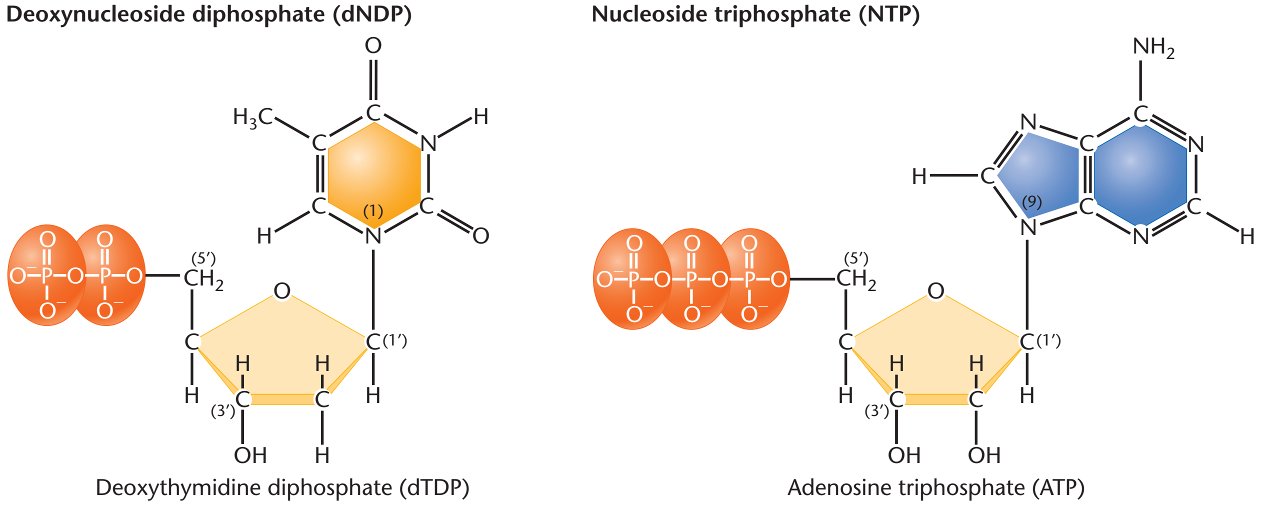

Deoxynucleoside diphosphate (dNDP)
Nucleoside triphosphate (NTP)
O
H_3C
C
N
H
(1)
(5′)
CH_2
C(1′)
(3′)
OH
Deoxythymidine diphosphate (dTDP)
NH_2
(9)
Adenosine triphosphate (ATP)

FIGURE 10.9 Structures of nucleoside diphosphates and triphosphates. Deoxythymidine diphosphate and adenosine triphosphate are diagrammed here.

energy produced is used to drive the reactions. As a result, both ATP and GTP are involved in many cellular activities, including numerous genetic events.

Polynucleotides

The linkage between two mononucleotides consists of a phosphate group linked to two sugars. It is called a **phosphodiester bond** because phosphoric acid has been joined to two alcohols (the hydroxyl groups on the two sugars) by an ester linkage on both sides. Figure 10.10(a) shows the phosphodiester bond in DNA. The same bond is found in RNA. Each structure has a **C-5′ end** and a **C-3′ end**. Two joined nucleotides form a **dinucleotide**; three nucleotides, a **trinucleotide**; and so forth. Short chains consisting of up to approximately 30 nucleotides linked together are called **oligonucleotides**; longer chains are called **polynucleotides**.

Because drawing polynucleotide structures, as shown in Figure 10.10(a), is time consuming and complex, a schematic shorthand method has been devised [Figure 10.10(b)]. The nearly vertical lines represent the pentose sugar; the nitrogenous base is attached at the top, in the C-1′ position. A diagonal line with the P in the middle of it is attached to the C-3′ position of one sugar and the C-5′ position of the neighboring sugar; it represents the phosphodiester bond. Several modifications of this shorthand method are in use, and they can be understood in terms of these guidelines.

Long polynucleotide chains account for the large molecular weight of DNA and explain its most important property—storage of vast quantities of genetic information. If each nucleotide position in this long chain can be occupied by any one of four nucleotides, extraordinary variation is possible. For example, a polynucleotide only 1000 nucleotides in length can be arranged 4^{1000} different ways, each one different from all other possible sequences. This potential variation in molecular structure is essential if DNA is to store the vast amounts of chemical information necessary to direct cellular activities.

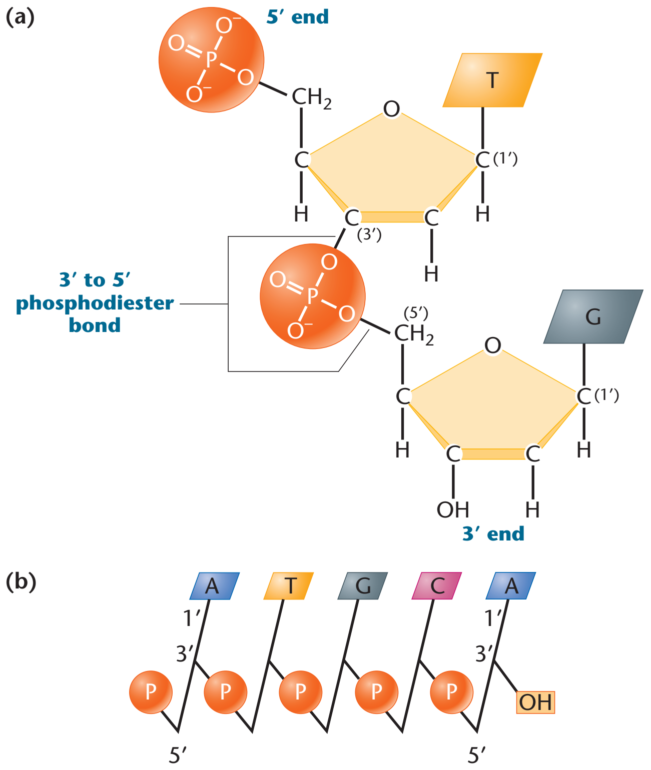


FIGURE 10.10 (a) Linkage of two nucleotides by the formation of a C-3′-to-C-5′ (3′-to-5′) phosphodiester bond, producing a dinucleotide. (b) Shorthand notation for a polynucleotide chain.

10.7 The Structure of DNA Holds the Key to Understanding Its Function

The previous sections in this chapter have established that DNA is the genetic material in all organisms (with certain viruses being the exception) and have provided details as to the basic chemical components making up nucleic acids. What remained to be deciphered was the precise structure of DNA. That is, how are polynucleotide chains organized into DNA, which serves as the genetic material? Is DNA composed of a single chain or more than one? If the latter is the case, how do the chains relate chemically to one another? Do the chains branch? And more important, how does the structure of this molecule relate to the various genetic functions served by DNA (i.e., storage, expression, replication, and mutation)?

From 1940 to 1953, many scientists were interested in solving the structure of DNA. Among others, Erwin Chargaff, Maurice Wilkins, Rosalind Franklin, Linus Pauling, Francis Crick, and James Watson sought information that might answer what many consider to be the most significant and intriguing question in the history of biology: *How does DNA serve as the genetic basis for life*? The answer was believed to depend strongly on the chemical structure and organization of the DNA molecule, given the complex but orderly functions ascribed to it.

In 1953, James Watson and Francis Crick proposed that the structure of DNA is in the form of a double helix. Their model was described in a short paper published in the journal *Nature*. In a sense, this publication was the finish of a highly competitive scientific race. Watson's book *The Double Helix* recounts the human side of the scientific drama that eventually led to the elucidation of DNA structure.

The data available to Watson and Crick, crucial to the development of their proposal, came primarily from two sources: (1) base composition analysis of hydrolyzed samples of DNA and (2) X-ray diffraction studies of DNA. Watson and Crick's analytical success can be attributed to their focus on building a model that conformed to the existing data. If the correct solution to the structure of DNA is viewed as a puzzle, Watson and Crick, working at the Cavendish Laboratory in Cambridge, England, were the first to fit the pieces together successfully.

Base-Composition Studies

Between 1949 and 1953, Erwin Chargaff and his colleagues used chromatographic methods to separate the four bases in DNA samples from various organisms. Quantitative methods were then used to determine the amounts of the four bases from each source. **Table 10.3(a)** lists some of Chargaff's original data. Parts (b) and (c) of the table show more recently derived base-composition information that reinforces Chargaff's findings. As we shall see, Chargaff's data were critical to the success of Watson and Crick as they devised the double-helical model of DNA. On the basis of these data, the following conclusions may be drawn:

1. As shown in **Table 10.3(b)**, the amount of adenine residues is proportional to the amount of thymine residues

TABLE 10.3 DNA Base-Composition Data

(a) Chargaff's Data*	Molar Proportions[a]			
	1	2	3	4
Organism/Source	A	T	G	C
Ox thymus	26	25	21	16
Ox spleen	25	24	20	15
Yeast	24	25	14	13
Avian tubercle bacilli	12	11	28	26
Human sperm	29	31	18	18

(c) G + C Content in Several Organisms	
Organism	%G + C
Phage T2	36.0
Drosophila	45.0
Maize	49.1
Euglena	53.5
Neurospora	53.7

(b) Base Compositions of DNAs from Various Sources								
	Base Composition				Base Ratio		Combined Base Ratios	
	1	2	3	4	5	6	7	8
Organism	A	T	G	C	A/T	G/C	(A + G)/(C + T)	(A + T)/(C + G)
Human	30.9	29.4	19.9	19.8	1.05	1.00	1.04	1.52
Sea urchin	32.8	32.1	17.7	17.3	1.02	1.02	1.02	1.58
E. coli	24.7	23.6	26.0	25.7	1.04	1.01	1.03	0.93
Sarcina lutea	13.4	12.4	37.1	37.1	1.08	1.00	1.04	0.35
T7 bacteriophage	26.0	26.0	24.0	24.0	1.00	1.00	1.00	1.08

**Source:* Data from Chargaff, 1950.

[a] Moles of nitrogenous constituent per mole of P. (Often, the recovery was less than 100 percent.)

in DNA (columns 1, 2, and 5). Also, the amount of guanine residues is proportional to the amount of cytosine residues (columns 3, 4, and 6).

2. Based on this proportionality, the sum of the purines (A + G) equals the sum of the pyrimidines (C + T) as shown in column 7.
3. The percentage of (G + C) does not necessarily equal the percentage of (A + T). As you can see, this ratio varies greatly among organisms, as shown in column 8 and in part (c) of Table 10.3.

These conclusions indicate a definite pattern of base composition in DNA molecules. The data provided the initial clue to the DNA puzzle. In addition, they directly refute Levene's tetranucleotide hypothesis, which stated that all four bases are present in equal amounts.

X-Ray Diffraction Analysis

When fibers of a DNA molecule are subjected to X-ray bombardment, the X rays scatter (diffract) in a pattern that depends on the molecule's atomic structure. The pattern of diffraction can be captured as spots on photographic film and analyzed for clues to the overall shape of and regularities within the molecule. This process, **X-ray diffraction analysis,** was applied successfully to the study of protein structure by Linus Pauling and other chemists. The technique had been attempted on DNA as early as 1938 by William Astbury. By 1947, he had detected a periodicity of 3.4-angstrom unit (3.4-Å) repetitions* within the structure of the molecule, which suggested to him that the bases were stacked like coins on top of one another.

Between 1950 and 1953, Rosalind Franklin, working in the laboratory of Maurice Wilkins, obtained improved X-ray data from more purified samples of DNA (Figure 10.11). Her work confirmed the 3.4-Å periodicity seen by Astbury and suggested that the structure of DNA was some sort of helix. However, she did not propose a definitive model. Pauling had analyzed the work of Astbury and others and incorrectly proposed that DNA was a triple helix.

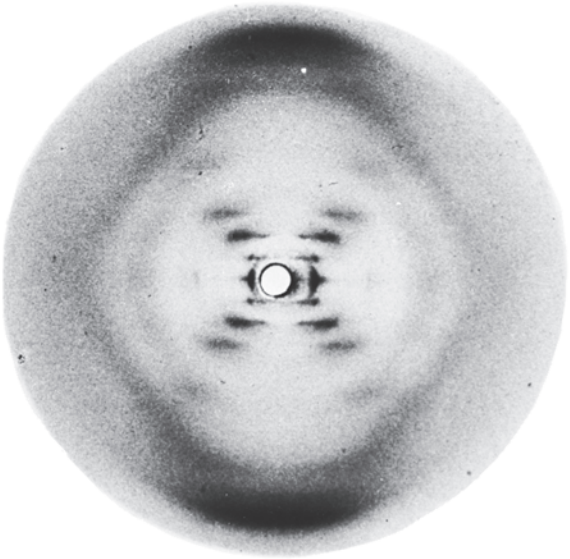


FIGURE 10.11 X-ray diffraction photograph by Rosalind Franklin using the B form of purified DNA fibers. The strong arcs on the periphery represent closely spaced aspects of the molecule, allowing scientists to estimate the periodicity of the nitrogenous bases, which are 3.4 Å apart. The inner cross pattern of spots reveals the grosser aspects of the molecule, indicating its helical nature.

The Watson–Crick Model

Watson and Crick published their analysis of DNA structure in 1953. By building models based on the above-mentioned parameters, they arrived at the double-helical form of DNA shown in Figure 10.12(a) on p. 227.

This model has the following major features:

1. Two long polynucleotide chains are coiled around a central axis, forming a right-handed double helix.
2. The two chains are **antiparallel;** that is, their C-5′-to-C-3′ orientations run in opposite directions.
3. The bases of both chains are flat structures lying perpendicular to the axis; they are "stacked" on one another, 3.4 Å (0.34 nm) apart, on the inside of the double helix.
4. The nitrogenous bases of opposite chains are *paired* as the result of the formation of hydrogen bonds; in DNA, only A-T and C-G pairs occur.
5. Each complete turn of the helix is 34 Å (3.4 nm) long; thus, each turn of the helix is the length of a series of 10 base pairs.
6. A larger **major groove** alternating with a smaller **minor groove** winds along the length of the molecule.
7. The double helix has a diameter of 20 Å (2.0 nm).

The nature of the base pairing (point 4 above) is the model's most significant feature in terms of explaining its genetic functions. Before we discuss it, however, several other important features warrant emphasis. First, the antiparallel arrangement of the two chains is a key part of the double-helix model. While one chain runs in the 5′-to-3′ orientation (what seems right side up to us), the other chain goes in the 3′-to-5′ orientation (and thus appears upside down). This is indicated in Figure 10.12(b) and (c). Given the bond angles in the structures of the various nucleotide components, the double helix could not be constructed easily if both chains ran parallel to one another.

Second, the right-handed nature of the helix modeled by Watson and Crick is best appreciated by comparison

*Today, measurement in nanometers (nm) is favored (1 nm = 10 Å).

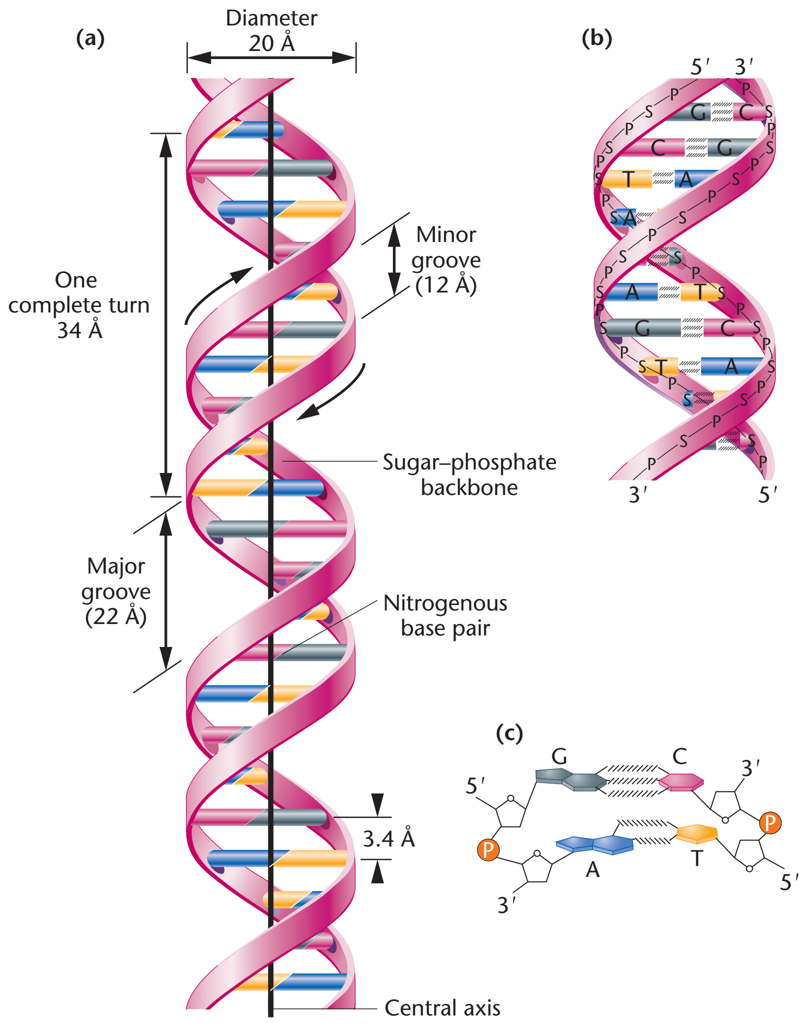


FIGURE 10.12 (a) The DNA double helix as proposed by Watson and Crick. The ribbon-like strands represent the sugar-phosphate backbones, and the horizontal rungs depict the nitrogenous base pairs, of which there are 10 per complete turn. The major and minor grooves are apparent. A solid vertical line shows the central axis. (b) A detailed view depicting the bases, sugars, phosphates, and hydrogen bonds of the helix. (c) A demonstration of the antiparallel arrangement of the chains and the horizontal stacking of the bases.

with its left-handed counterpart, which is a mirror image, as shown in **Figure 10.13**. The conformation in space of the right-handed helix is most consistent with the data that were available to Watson and Crick, although an alternative form of DNA (Z-DNA) does exist as a left-handed helix, as we will soon discuss in Section 10.8.

The key to the model proposed by Watson and Crick is the specificity of base pairing. Chargaff's data suggested that A was equal in amount to T and that G was equal in amount to C. Watson and Crick realized that pairing A with T and C with G would account for these proportions, and that such pairing could occur as a result of hydrogen bonds between base pairs [Figure 10.12(b)], which would also provide the chemical stability necessary to hold the two chains together. Arrangement of the components in this way produces the major and minor grooves along the molecule's length. Furthermore, a purine (A or G) opposite a pyrimidine (T or C) on each "rung of the spiral staircase" in the proposed helix accounts for the 20-Å (2-nm) diameter suggested by X-ray diffraction studies.

The specific A-T and C-G base pairing is described as **complementarity** and results from the chemical affinity that produces the hydrogen bonds in each pair of bases. As we will see, complementarity is very important in the processes of DNA replication and gene expression.

Two questions are particularly worthy of discussion. First, why aren't other base pairs possible? Watson and Crick discounted the pairing of A with G or of C with T because these would represent purine—purine and pyrimidine—pyrimidine pairings, respectively. Such pairings would lead to aberrant diameters of, in one case, more than and, in the other case, less than 20 Å because of the respective sizes of the purine and pyrimidine rings. In addition, the three-dimensional configurations that would be formed by such pairings would not produce an alignment that allows sufficient hydrogen-bond formation. It is for this reason that A-C and G-T pairings were also discounted, even though those pairs would each consist of one purine and one pyrimidine.

The second question concerns the properties of hydrogen bonds. Just what is the nature of such a bond, and is it strong enough to stabilize the helix? A **hydrogen bond** is a very weak electrostatic attraction between

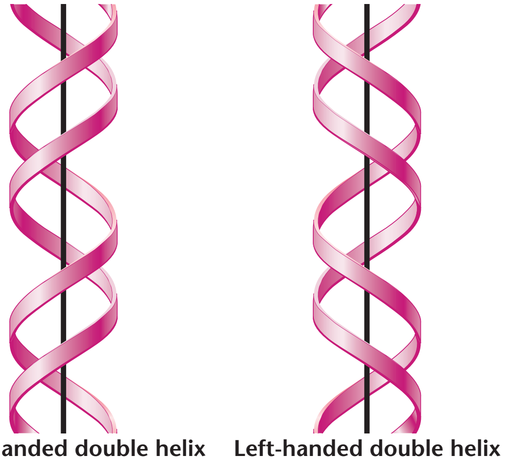


FIGURE 10.13 The right- and left-handed helical forms of DNA. Note that they are mirror images of one another.

a covalently bonded hydrogen atom and an atom with an unshared electron pair. The hydrogen atom assumes a partial positive charge, while the unshared electron pair—characteristic of covalently bonded oxygen and nitrogen atoms—assumes a partial negative charge. These opposite charges are responsible for the weak chemical attraction that is the basis of the hydrogen bond. As oriented in the double helix, adenine forms two hydrogen bonds with thymine, and guanine forms three hydrogen bonds with cytosine (**Figure 10.14**). Although two or three hydrogen bonds taken alone are energetically very weak, thousands of bonds in tandem (as found in long polynucleotide chains) provide great stability to the helix.

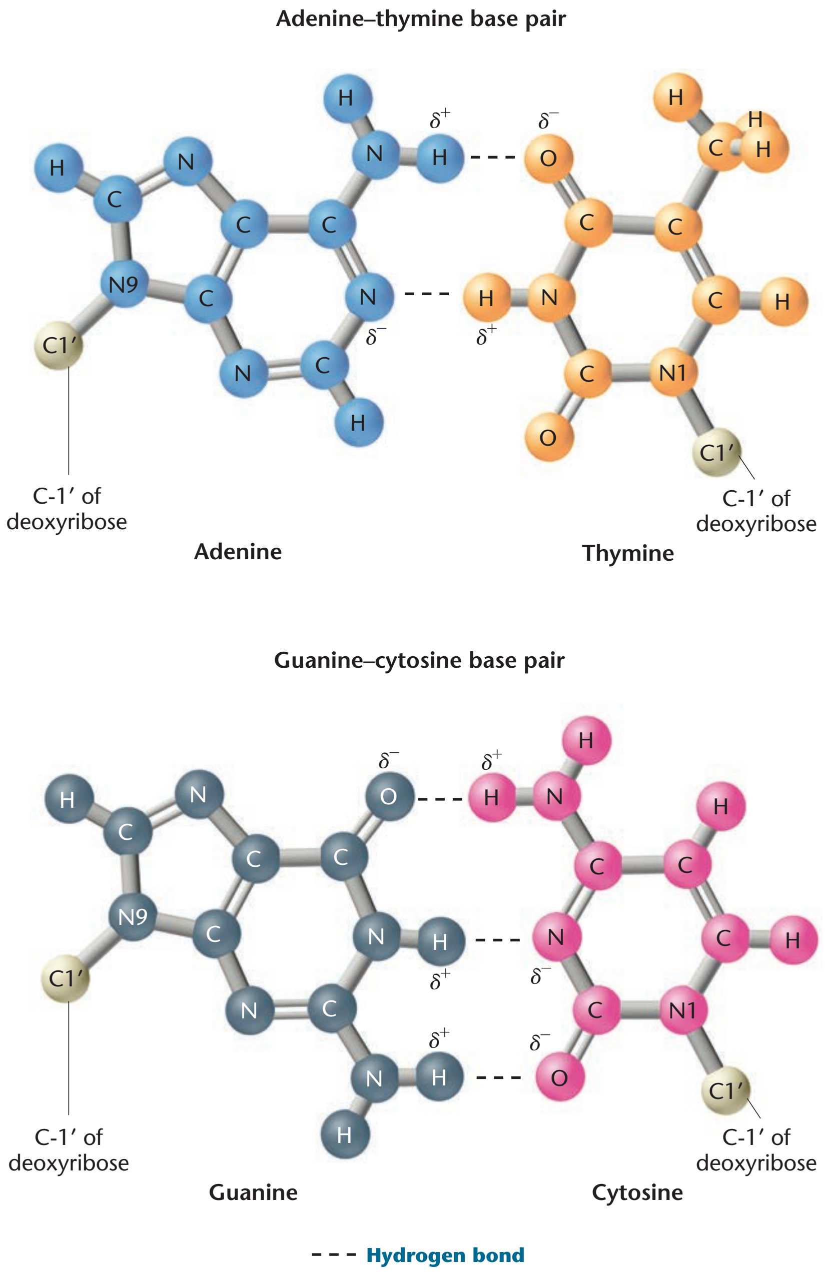


FIGURE 10.14 Ball-and-stick models of A-T and G-C base pairs. The dashes (---) represent the hydrogen bonds that form between bases.

Another stabilizing factor is the arrangement of sugars and bases along the axis. In the Watson–Crick model, the hydrophobic ("water-fearing") nitrogenous bases are stacked almost horizontally on the interior of the axis and are thus shielded from the watery environment that surrounds the molecule within the cell. The hydrophilic ("water-loving") sugar-phosphate backbones are on the outside of the axis, where both components may interact with water. These molecular arrangements provide significant chemical stabilization to the helix.

A more recent and accurate analysis of the form of DNA that served as the basis for the Watson–Crick model has revealed a minor structural difference between the substance and the model. A precise measurement of the number of base pairs per turn has demonstrated a value of 10.4, rather than the 10.0 predicted by Watson and Crick. In the classic model, each base pair is rotated 36° around the helical axis relative to the adjacent base pair, but the new finding requires a rotation of 34.6°. This results in slightly more than 10 base pairs per 360° turn.

The Watson–Crick model had an instant effect on the emerging discipline of molecular biology. Even in their initial 1953 article in *Nature*, the authors observed, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." Two months later, Watson and Crick pursued this idea in a second article in *Nature*, suggesting a specific mechanism of replication of DNA—*the semiconservative mode of replication* (described in Chapter 11). The second article also alluded to two new concepts: (1) the storage of genetic information in the sequence of the bases and (2) the mutations or genetic changes that would result from an alteration of bases. These ideas have received vast amounts of experimental support since 1953 and are now universally accepted.

Watson and Crick's synthesis of ideas was highly significant with regard to subsequent studies of genetics and biology. The nature of the gene and its role in genetic mechanisms could now be viewed and studied in biochemical terms. Recognition of their work, along with that of Wilkins, led to all three receiving the Nobel Prize in Physiology and Medicine in 1962. Unfortunately, Rosalind

Franklin had died in 1958 at the age of 37, making her contributions ineligible for consideration, since the award is not given posthumously. The Nobel Prize was to be one of many such awards bestowed for work in the field of molecular genetics.

EVOLVING CONCEPT OF THE GENE

Based on the model of DNA put forward by Watson and Crick in 1953, the gene was viewed for the first time in molecular terms as a sequence of nucleotides in a DNA helix that encodes genetic information. ■

10.8 Alternative Forms of DNA Exist

Under different conditions of isolation, different conformations of DNA are seen. At the time when Watson and Crick performed their analysis, two forms—**A-DNA** and **B-DNA**—were known. Watson and Crick's analysis was based on Rosalind Franklin's X-ray studies of the B form, which is seen under aqueous, low-salt conditions and is believed to be the biologically significant conformation.

While DNA studies around 1950 relied on the use of X-ray diffraction, more recent investigations have been performed using **single-crystal X-ray analysis**. The earlier studies achieved resolution of about 5 Å, but single crystals diffract X rays at about 1 Å intervals, near atomic resolution. As a result, every atom is "visible," and much greater structural detail is available.

With this modern technique, A-DNA, which is prevalent under high-salt or dehydration conditions, has now been scrutinized. In comparison to B-DNA, A-DNA is slightly more compact, with nine base pairs in each complete turn of the helix, which is 23 Å (2.3 nm) in diameter (**Figure 10.15**). While it is also a right-handed helix, the orientation of the bases is somewhat different—they are tilted and displaced laterally in relation to the axis of the helix. As a result, the appearance of the major and minor grooves is modified. It seems doubtful that A-DNA occurs *in vivo* (under physiological conditions).

Still other forms of DNA right-handed helices have been discovered when investigated under various laboratory conditions. These have been designated C-, D-, E-, and most recently P-DNA. **C-DNA** is found under even greater dehydration conditions than those observed during the isolation of A- and B-DNA. Two other forms, **D-DNA** and **E-DNA**, occur in helices lacking guanine in their base composition. And most recently, it has been observed that if DNA is artificially stretched, still another conformation is assumed, called **P-DNA** (named for Linus Pauling).

Finally, still another form of DNA, called **Z-DNA**, was discovered in 1979, when a small synthetic DNA oligonucleotide containing only G-C base pairs was studied. Z-DNA takes on the rather remarkable configuration of a *left-handed double helix*. Speculation has occasionally run high over the possibility that regions of Z-DNA exist in the chromosomes of living organisms. The unique helical arrangement has the potential to provide an important recognition site for interaction with DNA-binding molecules. However, the extent to which Z-DNA, or any of the forms mentioned above, occur *in vivo* is still not clear.

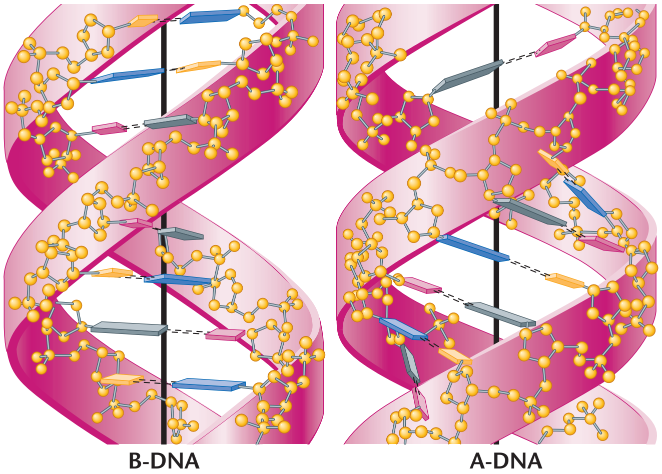


FIGURE 10.15 An artist's depiction illustrating the orientation of the base pairs of B-DNA and A-DNA. Note that in B-DNA the base pairs are perpendicular to the helix, while they are tilted and pulled away from the helix in A-DNA. In comparison to B-DNA, A-DNA also displays a slightly greater diameter and contains one more base pair per turn.

NOW SOLVE THIS

10.2 In sea urchin DNA, which is double stranded, 17.5 percent of the bases were shown to be cytosine (C). What percentages of the other three bases are expected to be present in this DNA?

■ **HINT:** *This problem asks you to extrapolate from one measurement involving a unique DNA molecule to three other values characterizing the molecule. The key to its solution is to understand the base-pairing rules in the Watson–Crick model of DNA.*

10.9 The Structure of RNA Is Chemically Similar to DNA, but Single Stranded

The structure of RNA molecules resembles DNA, with several important exceptions. Although RNA also has nucleotides linked into polynucleotide chains, the sugar ribose replaces deoxyribose, and the nitrogenous base uracil replaces thymine. Another important difference is that most RNA is single stranded, but there are important exceptions. RNA molecules often fold back on themselves to form double-stranded regions of complementary base pairs. In addition, some animal viruses that have RNA as their genetic material contain double-stranded helices. More recently, we have learned that double-stranded RNA molecules can regulate gene expression in eukaryotes (see Chapter 18).

As established earlier (see Figure 10.1), three major classes of cellular RNA molecules function during the expression of genetic information: **ribosomal RNA (rRNA), messenger RNA (mRNA),** and **transfer RNA (tRNA).** These molecules all originate as complementary copies of one of the two strands of DNA segments during the process of transcription. That is, their nucleotide sequence is complementary to the deoxyribonucleotide sequence of DNA that served as the template for their synthesis. Because uracil replaces thymine in RNA, uracil is complementary to adenine during transcription and during RNA base pairing.

Table 10.4 characterizes these major forms of RNA as found in bacteria and eukaryotic cells. Different RNAs are distinguished according to their sedimentation behavior in a centrifugal field and by their size (the number of nucleotides each contains). Sedimentation behavior depends on a molecule's density, mass, and shape, and its measure is called the **Svedberg coefficient (*S*).** While higher *S* values almost always designate molecules of greater molecular weight, the correlation is not direct; that is, a twofold increase in molecular weight does not lead to a twofold increase in *S*. This is because, in addition to a molecule's mass, the size and shape of the molecule also affect its rate of sedimentation (*S*). As you can see in Table 10.4, RNA molecules come in a wide range of sizes.

Ribosomal RNA usually constitutes about 80 percent of all RNA in an *E. coli* cell. Ribosomal RNAs are important structural components of **ribosomes**, which function as nonspecific workbenches where proteins are synthesized during translation. The various forms of rRNA found in bacteria and eukaryotes differ distinctly in size.

Messenger RNA molecules carry genetic information from the DNA of the gene to the ribosome. The mRNA molecules vary considerably in size, reflecting the range in the sizes of the proteins encoded by the mRNA as well as the different sizes of the genes serving as the templates for transcription of mRNA. While Table 10.4 shows that about 5 percent of RNA is mRNA in *E. coli*, this percentage varies from cell to cell and even at different times in the life of the same cell.

Transfer RNA, accounting for up to 15 percent of the RNA in a typical cell, is the smallest class (in terms of average molecule size) of these RNA molecules and carries amino acids to the ribosome during translation. Because more than one tRNA molecule interacts simultaneously with the ribosome, the molecule's smaller size facilitates these interactions.

We will discuss more detailed functions of these three classes of RNA later (see Chapters 13 and 14). While these RNAs represent the major forms of the molecule involved in genetic expression, other unique RNAs exist that perform various roles, especially in eukaryotes. For example, *telomerase RNA* and *RNA primers* are involved in DNA replication (Chapter 11), *small nuclear RNA (snRNA)* participates in processing mRNAs (Chapter 13), and *microRNA (miRNA)*, *short interfering RNA (siRNA)*, and *long noncoding RNA (lncRNA)* are involved in gene regulation (Chapter 18).

TABLE 10.4 Principal Classes of RNA

Class	% of Total RNA*	Svedberg Coefficient	Eukaryotic (E) or Prokaryotic (P)	Number of Nucleotides
Ribosomal (rRNA)	80	5	P and E	120
		5.8	E	160
		16	P	1542
		18	E	1874
		23	P	2904
		28	E	4718
Transfer (tRNA)	15	4	P and E	75–90
Messenger (mRNA)	5	varies	P and E	100–10,000

*In *E. coli*.

NOW SOLVE THIS

10.3 German measles results from an infection of the rubella virus, which can cause a multitude of health problems in newborns. What conclusions can you reach from a viral nucleic acid analysis that reveals an A + G/U + C ratio of 1.13?

- **HINT:** *This problem asks you to analyze information about the chemical composition of a nucleic acid serving as the genetic material of a virus. The key to its solution is to apply your knowledge of nucleic acid chemistry, in particular your understanding of base-pairing.*

For more practice, see Problems 31 and 32.

10.10 Many Analytical Techniques Have Been Useful during the Investigation of DNA and RNA

Since 1953, the role of DNA as the genetic material and the role of RNA in transcription and translation have been clarified through detailed analysis of nucleic acids. Several important methods of analysis are based on the unique nature of the hydrogen bond that is so integral to the structure of nucleic acids. For example, if DNA is subjected to heat, the double helix is denatured and unwinds. During unwinding, the viscosity of DNA decreases and UV absorption increases (called the **hyperchromic shift**). A melting profile, in which OD_{260} is plotted against temperature, is shown for two DNA molecules in Figure 10.16. The midpoint of each curve is called the **melting temperature (T_m)** where 50 percent of the strands have unwound. The molecule with a higher T_m has a higher percentage of G-C base pairs than A-T base pairs since G-C pairs share three hydrogen bonds compared to the two bonds between A-T pairs.

The denaturation/renaturation of nucleic acids is the basis for one of the most useful techniques in molecular genetics—**molecular hybridization.** Provided that a reasonable degree of base complementarity exists between any two nucleic acid strands, denaturation can be reversed whereby molecular hybridization is possible. Duplexes can be re-formed between DNA strands, even from different organisms, and between DNA and RNA strands. For example, an RNA molecule will hybridize with the segment of DNA from which it was transcribed. As a result, nucleic acid **probes** are often used to identify complementary sequences.

The technique can even be performed using the DNA present in chromosomal preparations as the "target" for hybrid formation. This process is called ***in situ* molecular hybridization.** Mitotic cells are first fixed to slides and then subjected to hybridization conditions. Single-stranded DNA or RNA is added (a probe), and hybridization is monitored. The nucleic acid that is added may be either radioactive or contain a fluorescent label to allow its detection. In the former case, autoradiography is used.

Figure 10.17 illustrates the use of a fluorescent label. A short fragment of DNA that is complementary to DNA in the chromosomes' centromere regions has been hybridized. Fluorescence occurs only in the centromere regions and thus identifies each one along its chromosome. Because fluorescence is used, the technique is known by the acronym **FISH (fluorescent *in situ* hybridization).** The use of this technique to identify chromosomal locations housing specific genetic information has been a valuable addition to geneticists' repertoire of experimental techniques.

Electrophoresis

Another technique essential to the analysis of nucleic acids is **electrophoresis.** This technique may be adapted to separate different-sized fragments of DNA and RNA chains and

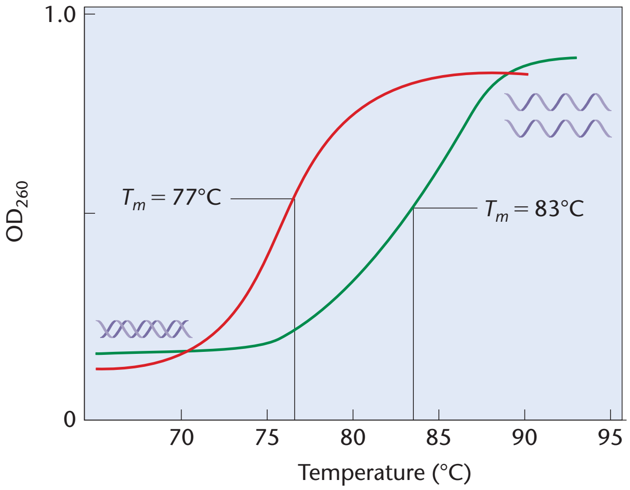


FIGURE 10.16 A melting profile shows the increase in UV absorption versus temperature (the hyperchromic effect) for two DNA molecules with different G-C contents. The molecule with a melting point (T_m) of 83°C has a greater G-C content than the molecule with a T_m of 77°C.

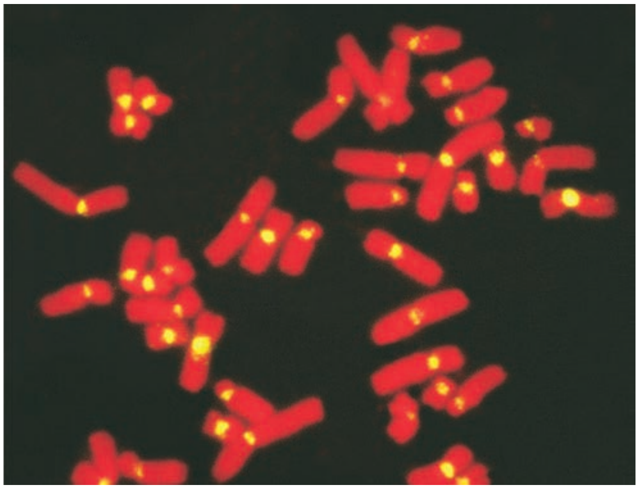


FIGURE 10.17 Fluorescent *in situ* hybridization (FISH) of human metaphase chromosomes. The probe, specific to centromeric DNA, produces a yellow fluorescence signal indicating hybridization. The red fluorescence is produced by propidium iodide counterstaining of chromosomal DNA.

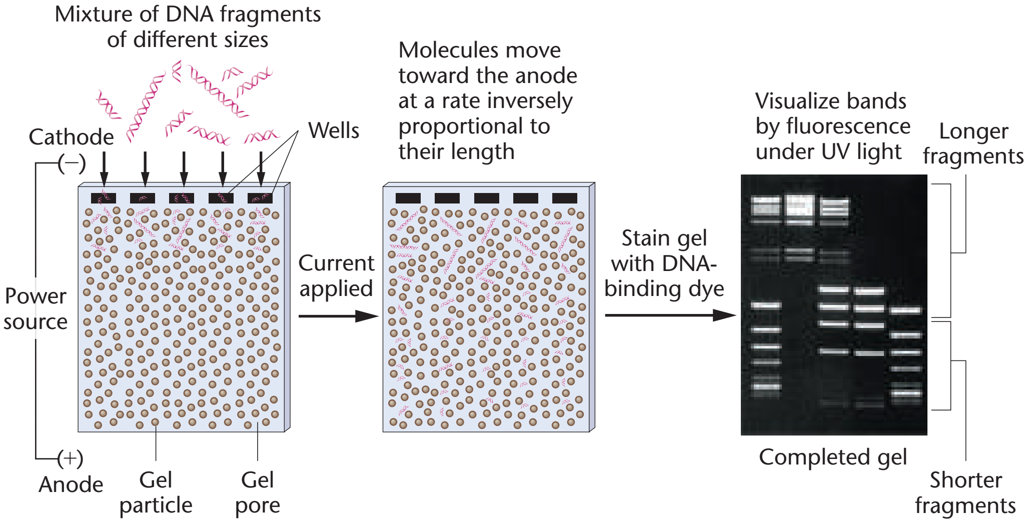
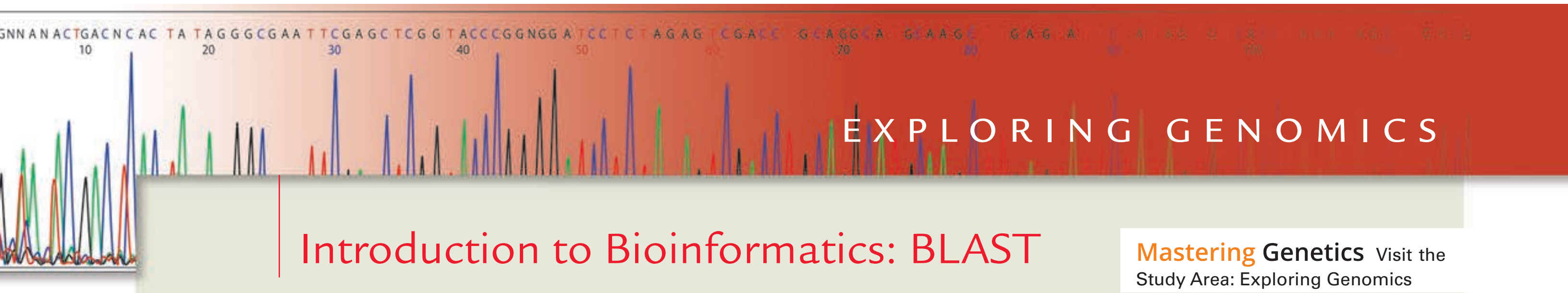

Mixture of DNA fragments of different sizes
Cathode
(−)
Wells
Power source
(+)
Anode
Gel particle
Gel pore
Current applied
Molecules move toward the anode at a rate inversely proportional to their length
Stain gel with DNA-binding dye
Visualize bands by fluorescence under UV light
Longer fragments
Completed gel
Shorter fragments

FIGURE 10.18 Electrophoretic separation of a mixture of DNA fragments that vary in length. The photograph at the right shows an agarose gel that reveals DNA bands stained by ethidium bromide.

is invaluable in current research investigations in molecular genetics.

Electrophoresis separates the molecules in a mixture by causing them to migrate under the influence of an electric field. A DNA sample is placed into a semisolid gel immersed in a solution that conducts electricity. Mixtures of molecules with a similar charge-to-mass ratio but of different sizes will migrate at different rates through the gel based on their size. For example, two polynucleotide chains of different lengths, such as 10 versus 20 nucleotides, are both negatively charged (based on the phosphate groups of the nucleotides) and will both move to the positively charged pole (the anode), but at different rates. Using a medium such as an **agarose gel,** which can be prepared with various pore sizes, the *shorter chains migrate at a faster rate through the gel than larger chains* (Figure 10.18). Once electrophoresis is complete, bands representing the variously sized molecules are identified by use of a fluorescent dye that binds to nucleic acids. The resolving power of some gels is so great that polynucleotides that vary by just one nucleotide in length may be separated.

Electrophoretic separation of nucleic acids is at the heart of a variety of other commonly used research techniques. Of particular note are the various "blotting" techniques (e.g., Southern blots and Northern blots). (We will discuss these in detail later in Chapters 20 and 21.)



EXPLORING GENOMICS

Introduction to Bioinformatics: BLAST

Mastering Genetics Visit the Study Area: Exploring Genomics

In this chapter, we focused on the structural details of DNA. Later, you will learn how scientists can clone and sequence DNA (see Chapter 20). The explosion of DNA and protein sequence data that has occurred in the last 20 years has launched the field of *bioinformatics*, an interdisciplinary science that applies mathematics and computing technology to develop hardware and software for storing, sharing, comparing, and analyzing nucleic acid and protein sequence data.

A large number of sequence databases that make use of bioinformatics have been developed. An example is **GenBank** (http://www.ncbi.nlm.nih.gov/genbank/), which is the National Institutes of Health database of all publicly available sequence data. This global resource, with access to databases in Europe and Japan, currently contains more than 220 billion base pairs of sequence data!

Earlier (in the Exploring Genomics exercises for Chapter 5), you were introduced to the National Center for Biotechnology Information (NCBI) Genes and Disease site. Now we will use an NCBI application called **BLAST, Basic Local Alignment Search Tool**. BLAST is an invaluable program for searching through GenBank and other databases to find DNA- and protein-sequence similarities between cloned substances. It has many additional functions that we will explore in other exercises.

Exercise I – Introduction to BLAST

1. Access BLAST from the NCBI Web site at http://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. Click on "nucleotide blast." This feature allows you to search DNA databases to look for a similarity between a sequence you enter and other sequences in the database. Do a nucleotide search with the following sequence:

   CCAGAGTCCAGCTGCTGCTCATA
   CTACTGATACTGCTGGG
3. Imagine that this sequence is a short part of a gene you cloned in your laboratory. You want to know if this gene or others with similar sequences have been discovered. Enter this sequence into the "Enter Query Sequence" text box at the top of the page. Near the bottom of the page, under the "Program Selection" category, choose "blastn"; then click on the "BLAST" button at the bottom of the page to run the search. It may take several minutes for results to be available because BLAST is using powerful algorithms to scroll through billions of bases of sequence data! A new page will appear with the results of your search.
4. On the search results page, below the Graphic Summary you will see a category called Descriptions and a table showing significant matches to the sequence you searched with (called the query sequence). BLAST determines significant matches based on statistical measures that consider the length of the query sequence, the number of matches with sequences in the database, and other factors. Significant *alignments*, regions of significant similarity in the query and subject sequences, typically have E values less than 1.0.
5. The top part of the table lists matches to transcripts (mRNA sequences), and the lower part lists matches to genomic DNA sequences, in order of highest to lowest number of matches.
6. Alignments are indicated by horizontal lines. BLAST adjusts for gaps in the sequences, that is, for areas that may not align precisely because of missing bases in otherwise similar sequences. Scroll below the table to see the aligned sequences from this search, and then answer the following questions:

   a. What were the top three matches to your query sequence?

   b. For each alignment, BLAST also indicates the percent *identity* and the number of gaps in the match between the query and subject sequences. What was the percent identity for the top three matches? What percentage of each aligned sequence showed gaps indicating sequence differences?

   c. Click on the links for the first matched sequence (far-right column). These will take you to a wealth of information, including the size of the sequence; the species it was derived from; a PubMed-linked chronology of research publications pertaining to this sequence; the complete sequence; and if the sequence encodes a polypeptide, the predicted amino acid sequence coded by the gene. Skim through the information presented for this gene. What is the gene's function?
7. A BLAST search can also be done by entering the *accession number* for a sequence, which is a unique identifying number assigned to a sequence before it can be put into a database. For example, search with the accession number NM_007305. What did you find?
8. Run a BLAST search using the sequences or accession numbers listed in a. to c. below. In each case, after entering the accession number or sequence in the "Enter Query Sequence" box, go to the "Choose Search Set" box and click on the "Others" button for database. Then go to the "Program Selection" box and click "megablast" before running your search. These features will allow you to align the query sequence with similar genes from a number of other species. When each search is completed, explore the information BLAST provides so that you can identify and learn about the gene encoded by the sequence.

   a. NM_001006650. What is the top sequence that aligns with the query sequence of this accession number and shows 100 percent sequence identity?

   b. DQ991619. What gene is encoded by this sequence?

   c. NC_007596. What living animal has a sequence similar to this one?

CASE STUDY Credit where credit is due

In the early 1950s, it became clear to many researchers that DNA was the cellular molecule that carries genetic information. However, an understanding of the genetic properties of DNA could only be achieved through a detailed knowledge of its structure. To this end, several laboratories began a highly competitive race to discover the three-dimensional structure of DNA, which ended when Watson and Crick published their now classic paper in 1953. Their model was based, in part, on an X-ray diffraction photograph of DNA taken by Rosalind Franklin (Figure 10.11). Two ethical issues surround this photo. First, the photo was given to Watson and Crick by Franklin's co-worker, Maurice Wilkins, without her knowledge or consent. Second, in their paper, Watson and Crick did not credit Franklin's contribution. The fallout from these lapses lasted for decades and raises some basic questions about ethics in science.

1. What vital clues were provided by Franklin's work to Watson and Crick about the molecular structure of DNA?

2. Was it ethical for Wilkins to show Franklin's unpublished photo to Watson and Crick without Franklin's consent? Would it have been more ethical for Watson and Crick to have offered Franklin co-authorship on this paper?

3. Given that these studies were conducted in the 1950s, how might gender have played a role in the fact that Rosalind Franklin did not receive due credit for her X-ray diffraction work?

See the Understanding Science: How Science *Really* Works Web site: "Credit and debt" (http://undsci.berkeley.edu/article/0_0_0/dna_13).

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Although both proteins and nucleic acids were initially considered as possible candidates for genetic material, proteins were initially favored.
2. By 1952, transformation studies and experiments using bacteria infected with bacteriophages strongly suggested that DNA is the genetic material in bacteria and most viruses.
3. Although initially only indirect observations supported the hypothesis that DNA controls inheritance in eukaryotes, subsequent studies involving recombinant DNA techniques and transgenic mice provided direct experimental evidence that the eukaryotic genetic material is DNA.
4. RNA serves as the genetic material in some bacteriophages as well as some plant and animal viruses.
5. As proposed by Watson and Crick, DNA exists in the form of a right-handed double helix composed of two long antiparallel polynucleotide chains held together by hydrogen bonds formed between complementary, nitrogenous base pairs.
6. The second category of nucleic acids important in genetic function is RNA, which is similar to DNA with the exceptions that it is usually single stranded, the sugar ribose replaces the deoxyribose, and the pyrimidine uracil replaces thymine.
7. Various methods of analysis of nucleic acids, particularly molecular hybridization and electrophoresis, have led to studies essential to our understanding of genetic mechanisms.

INSIGHTS AND SOLUTIONS

The current chapter, in contrast to previous discussions, does not emphasize genetic problem solving. Instead, it recounts some of the initial experimental analyses that launched the era of molecular genetics. Accordingly, our "Insights and Solutions" section shifts its emphasis to experimental rationale and analytical thinking, an approach that will continue to be used whenever appropriate.

1. (a) Based strictly on your scrutiny of the transformation data of Avery, MacLeod, and McCarty, what objection might be made to the conclusion that DNA is the genetic material? What other conclusion might be considered?

   (b) What observations, including later ones, argue against this objection?

   Solution:

   (a) Based solely on their results, it may be concluded that DNA is essential for transformation. However, DNA might have been a substance that caused capsular formation by *directly* converting nonencapsulated cells to cells with a capsule. That is, DNA may simply have played a catalytic role in capsular synthesis, leading to cells displaying smooth type III colonies.

   (b) First, transformed cells pass the trait onto their progeny cells, thus supporting the conclusion that DNA is responsible for heredity, not for the direct production of polysaccharide coats. Second, subsequent transformation studies over a period of five years showed that other traits, such as antibiotic resistance, could be transformed. Therefore, the transforming factor has a broad general effect, not one specific to polysaccharide synthesis. This observation is more in keeping with the conclusion that DNA is the genetic material.

2. If RNA were the universal genetic material, how would it have affected the Avery experiment and the Hershey—Chase experiment?

   Solution: In the Avery experiment, digestion of the soluble filtrate with RNase, rather than DNase, would have eliminated transformation. Had this occurred, Avery and his colleagues would have concluded that RNA was the transforming factor. Hershey and Chase would have obtained identical results, since ^{32}P would also label RNA but not protein. Had they been using a bacteriophage with RNA as its nucleic acid, and had they known this, they would have concluded that RNA was responsible for directing the reproduction of their bacteriophage.

3. A quest to isolate an important disease-causing organism was successful, and molecular biologists were hard at work analyzing the results. The organism contained as its genetic material a remarkable nucleic acid with a base composition of A = 21 percent, C = 29 percent, G = 29 percent, U = 21 percent. When heated, it showed a major hyperchromic shift, and when the reassociation kinetics were studied, the nucleic acid of this organism reannealed more slowly than that of phage T4 and *E. coli*. T4 contains 10^5 nucleotide pairs.

Analyze this information carefully, and draw *all* possible conclusions about the genetic material of this organism, based strictly on the preceding observations. As a test of your model, make one prediction that if upheld would strengthen your hypothesis about the nature of this molecule.

Solution: First of all, because of the presence of uracil (U), the molecule appears to be RNA. In contrast to normal RNA, this one has base ratios of A/U = G/C = 1, suggesting that the molecule may be a double helix. The hyperchromic shift and reassociation kinetics support this hypothesis. In the kinetic study, since none of the nucleic acid segments reannealed more rapidly than bacterial or viral nucleic acid, there is no repetitive sequence RNA. Furthermore, the total length of unique-sequence DNA is greater than that of either phage T4 (10^5 nucleotide pairs) or *E. coli*. A prediction might be made concerning the sugars. Our model suggests that ribose rather than deoxyribose should be present. If so, this observation would support the hypothesis that RNA is the genetic material in this organism.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. HOW DO WE KNOW? In this chapter, we first focused on the information that showed DNA to be the genetic material and then discussed the structure of DNA as proposed by Watson and Crick. We concluded the chapter by describing various techniques developed to study DNA. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
(a) How were scientists able to determine that DNA, and not some other molecule, serves as the genetic material in bacteria and bacteriophages?
(b) How do we know that DNA also serves as the genetic material in eukaryotes such as humans?
(c) How was it determined that the structure of DNA is a double helix with the two strands held together by hydrogen bonds formed between complementary nitrogenous bases?
(d) How do we know that G pairs with C and that A pairs with T as complementary base pairs are formed?
2. CONCEPT QUESTION Review the Chapter Concepts list on p. 213 . Most center around DNA and RNA and their role of serving as the genetic material. Write a short essay that contrasts these molecules, including a comparison of advantages conferred by their structure that each of them has over the other in serving in this role.
3. Discuss the reasons proteins were generally favored over DNA as the genetic material before 1940. What was the role of the tetranucleotide hypothesis in this controversy?
4. Contrast the contributions made to an understanding of transformation by Griffith and by Avery and his colleagues.
5. When Avery and his colleagues had obtained what was concluded to be the transforming factor from the IIIS virulent cells, they treated the fraction with proteases, RNase, and DNase, followed in each case by the assay for retention or loss of transforming ability. What were the purpose and results of these experiments? What conclusions were drawn?
6. Why were ^{32}P and ^{35}S chosen for use in the Hershey–Chase experiment? Discuss the rationale and conclusions of this experiment.
7. Does the design of the Hershey–Chase experiment distinguish between DNA and RNA as the molecule serving as the genetic material? Why or why not?
8. What observations are consistent with the conclusion that DNA serves as the genetic material in eukaryotes? List and discuss them.
9. What are the exceptions to the general rule that DNA is the genetic material in all organisms? What evidence supports these exceptions?
10. Draw the chemical structure of the three components of a nucleotide, and then link the three together. What atoms are removed from the structures when the linkages are formed?
11. How are the carbon and nitrogen atoms of the sugars, purines, and pyrimidines numbered?
12. Adenine may also be named 6-amino purine. How would you name the other four nitrogenous bases, using this alternative system? (O is indicated by “oxy-,” and CH_3 by “methyl.”)
13. Draw the chemical structure of a dinucleotide composed of A and G. Opposite this structure, draw the dinucleotide composed of T and C in an antiparallel (or upside-down) fashion. Form the possible hydrogen bonds.
14. Describe the various characteristics of the Watson–Crick double-helix model for DNA.
15. What evidence did Watson and Crick have at their disposal in 1953? What was their approach in arriving at the structure of DNA?
16. What might Watson and Crick have concluded had Chargaff’s data from a single source indicated the following?

	A	T	G	C
%	29	19	21	31

Why would this conclusion be contradictory to Wilkins’s and Franklin’s data?

17. How do covalent bonds differ from hydrogen bonds? Define base complementarity.

18. List three main differences between DNA and RNA.

19. What are the three major types of RNA molecules? How is each related to the concept of information flow?

20. How is the absorption of ultraviolet light by DNA and RNA important in the analysis of nucleic acids?

21. What is the physical state of DNA after it is heated and denatured?

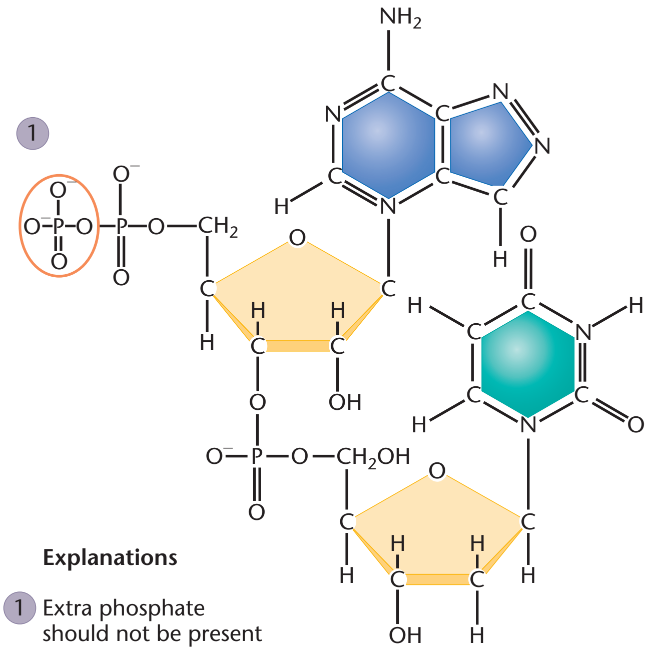
22. What is the hyperchromic effect? How is it measured? What does T_m imply?

23. Why is T_m related to base composition?

24. What is the chemical basis of molecular hybridization?

25. What did the Watson–Crick model suggest about the replication of DNA?

26. A genetics student was asked to draw the chemical structure of an adenine- and thymine-containing dinucleotide derived from DNA. The answer is shown here:

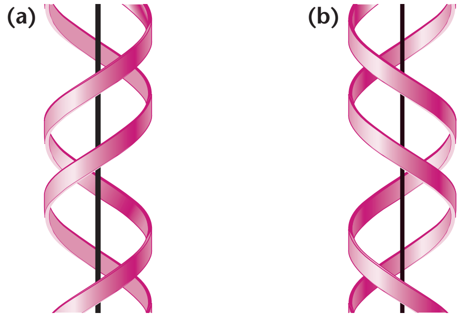


Explanations

1 Extra phosphate should not be present

The student made more than six major errors. One of them is circled, numbered 1, and explained. Find five others. Circle them, number them 2 through 6, and briefly explain each in the manner of the example given.

27. Considering the information in this chapter on B- and Z-DNA and right- and left-handed helices, carefully analyze structures (a) and (b) below and draw conclusions about their helical nature. Which is right handed and which is left handed?



28. One of the most common spontaneous lesions that occurs in DNA under physiological conditions is the hydrolysis of the amino group of cytosine, converting the cytosine to uracil. What would be the effect on DNA structure of a uracil group replacing cytosine?

29. In some organisms, cytosine is methylated at carbon 5 of the pyrimidine ring after it is incorporated into DNA. If a 5-methyl cytosine molecule is then hydrolyzed, as described in Problem 28, what base will be generated?

30. Because of its rapid turnaround time, fluorescent *in situ* hybridization (FISH) is commonly used in hospitals and laboratories as an aneuploid screen of cells retrieved from amniocentesis and chorionic villus sampling (CVS). Chromosomes 13, 18, 21, X, and Y (see Chapter 8) are typically screened for aneuploidy in this way. Explain how FISH might be accomplished using amniotic or CVS samples and why the above chromosomes have been chosen for screening.

Extra-Spicy Problems

31. A primitive eukaryote was discovered that displayed a unique nucleic acid as its genetic material. Analysis provided the following information:

(a) The general X-ray diffraction pattern is similar to that of DNA, but with somewhat different dimensions and more irregularity.

(b) A major hyperchromic shift is evident upon heating and monitoring UV absorption at 260 nm.

(c) Base-composition analysis reveals four bases in the following proportions:

Adenine	=	8%
Guanine	=	37%
Xanthine	=	37%
Hypoxanthine	=	18%

(d) About 75 percent of the sugars are deoxyribose, while 25 percent are ribose.

Postulate a model for the structure of this molecule that is consistent with the foregoing observations.

32. *Newsdate: March 1, 2030.* A unique creature has been discovered during exploration of outer space. Recently, its genetic material has been isolated and analyzed. This material is similar in some ways to DNA in its chemical makeup. It contains in abundance the 4-carbon sugar erythrose and a molar equivalent of phosphate groups. In addition, it contains six nitrogenous bases: adenine (A), guanine (G), thymine (T), cytosine (C), hypoxanthine (H), and xanthine (X). These bases exist in the following relative proportions:

$$A = T = H \text{ and } C = G = X$$

X-ray diffraction studies have established a regularity in the molecule and a constant diameter of about 30 Å. Together, these data have suggested a model for the structure of this molecule.

(a) Propose a general model of this molecule. Describe it briefly.

(b) What base-pairing properties must exist for H and for X in the model?

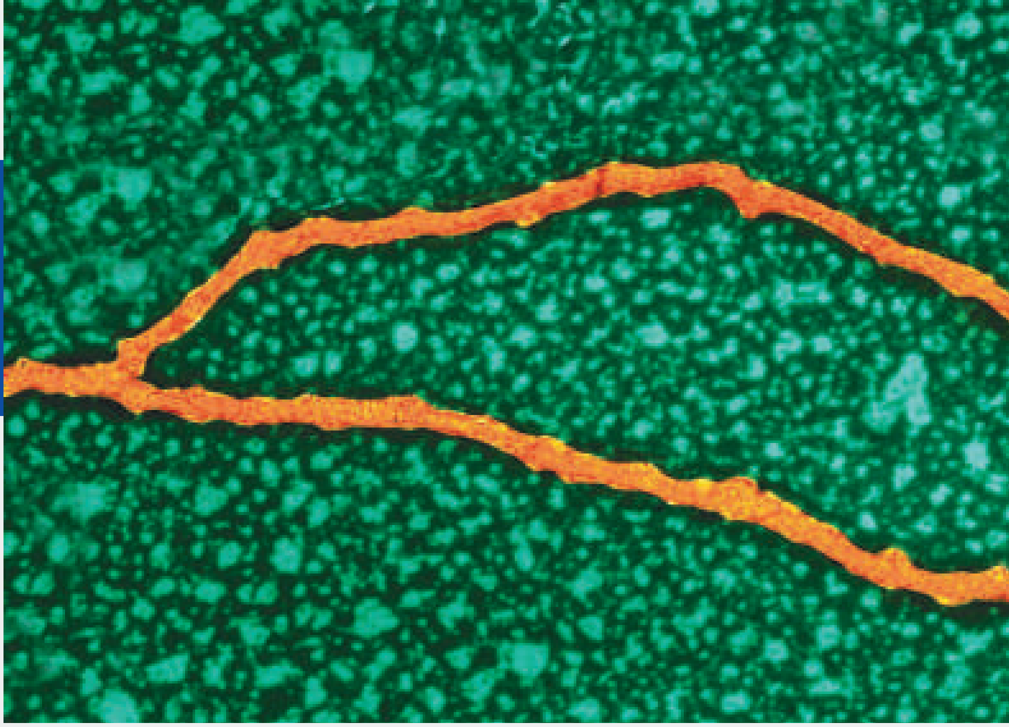
(c) Given the constant diameter of 30 Å, do you think that *either* (i) both H and X are purines or both pyrimidines, *or* (ii) one is a purine and one is a pyrimidine?

33. During gel electrophoresis, DNA molecules can easily be separated according to size because all DNA molecules have the same charge-to-mass ratio and the same shape (long rod). Would you expect RNA molecules to behave in the same manner as DNA during gel electrophoresis? Why or why not?

34. DNA and RNA are chemically very similar but are distinguished, in large part, by the presence of a 2'-OH group in RNA and a 2'-H group in DNA. Why do you suppose that both DNA and RNA have 3'-OH groups and we do not typically find nucleic acids within cells that have 3'-H groups?

35. Electrophoresis is an extremely useful procedure when applied to analysis of nucleic acids as it can resolve molecules of different sizes with relative ease and accuracy. Large molecules migrate more slowly than small molecules in agarose gels. However, the fact that nucleic acids of the same length may exist in a variety of conformations can often complicate the interpretation of electrophoretic separations. For instance, when a single species of a bacterial plasmid is isolated from cells, the individual plasmids may exist in three forms (depending on the genotype of their host and conditions of isolation): superhelical/supercoiled (form I), nicked/open circle (form II), and linear (form III). Form I is compact and very tightly coiled, with both DNA strands continuous. Form II exists as a loose circle because one of the two DNA strands has been broken, thus releasing the supercoil. All three have the same mass, but each will migrate at a different rate through a gel. Based on your understanding of gel composition and DNA migration, predict the relative rates of migration of the three DNA structures mentioned above.

11



DNA Replication and Recombination

Transmission electron micrograph of human DNA from a HeLa cell, illustrating a replication fork characteristic of active DNA replication.

CHAPTER CONCEPTS

- Genetic continuity between parental and progeny cells is maintained by semiconservative replication of DNA, as predicted by the Watson–Crick model.
- Semiconservative replication uses each strand of the parent double helix as a template, and each newly replicated double helix includes one "old" and one "new" strand of DNA.
- DNA synthesis is a complex but orderly process, occurring under the direction of a myriad of enzymes and other proteins.
- DNA synthesis involves the polymerization of nucleotides into polynucleotide chains.
- DNA synthesis is similar in bacteria and eukaryotes, but more complex in eukaryotes.
- In eukaryotes, DNA synthesis at the ends of chromosomes (telomeres) poses a special problem, overcome by a unique RNA-containing enzyme, telomerase.
- Genetic recombination is an important process leading to the exchange of genetic information between DNA molecules.

Following Watson and Crick's proposal for the structure of DNA, scientists focused their attention on how this molecule is replicated. Replication is an essential function of the genetic material and must be executed precisely if genetic continuity between cells is to be maintained following cell division. It is an enormous, complex task. Consider for a moment that more than 3×10^9 (3 billion) base pairs exist within the human genome. To duplicate faithfully the DNA of just one of these chromosomes requires a mechanism of extreme precision. Even an error rate of only 10^6 (one in a million) will still create 3000 errors (obviously an excessive number) during each replication cycle of the genome. Although it is not error free, and much of evolution would not have occurred if it were, an extremely accurate system of DNA replication has evolved in all organisms.

As Watson and Crick wrote at the end of their classic 1953 paper that announced the double helical model of DNA, "It has not escaped our notice that the specific pairing (A-T and C-G) we have postulated immediately suggests a copying mechanism for the genetic material." Called *semiconservative replication,* this mode of DNA duplication was soon to receive strong support from numerous studies of viruses, bacteria, and eukaryotes. Once the general *mode* of replication was clarified, research to determine the precise details of *DNA synthesis* intensified. What has since been discovered is that numerous enzymes and other proteins are needed to copy a DNA helix. Because of the complexity of the chemical events during synthesis, this subject remains an extremely active area of research.

In this chapter, we will discuss the general mode of replication, as well as the specific details of DNA synthesis. The research leading to such knowledge is another link in our understanding of life processes at the molecular level.

11.1 DNA Is Reproduced by Semiconservative Replication

Watson and Crick recognized that, because of the arrangement and nature of the nitrogenous bases, each strand of a DNA double helix could serve as a template for the synthesis of its complement (**Figure 11.1**). They proposed that, if the helix were unwound, each nucleotide along the two parent strands would have an affinity for its complementary nucleotide. As we learned earlier in the text (see Chapter 10), the complementarity is due to the potential hydrogen bonds that can be formed. If thymidylic acid (T) were present, it would "attract" adenylic acid (A); if guanidylic acid (G) were present, it would attract cytidylic acid (C); likewise, A would attract T, and C would attract G. If these nucleotides were then covalently linked into polynucleotide chains along both templates, the result would be the production of two identical double strands of DNA. Each replicated DNA molecule would consist of one "old" and one "new" strand, hence the reason for the name **semiconservative replication.**

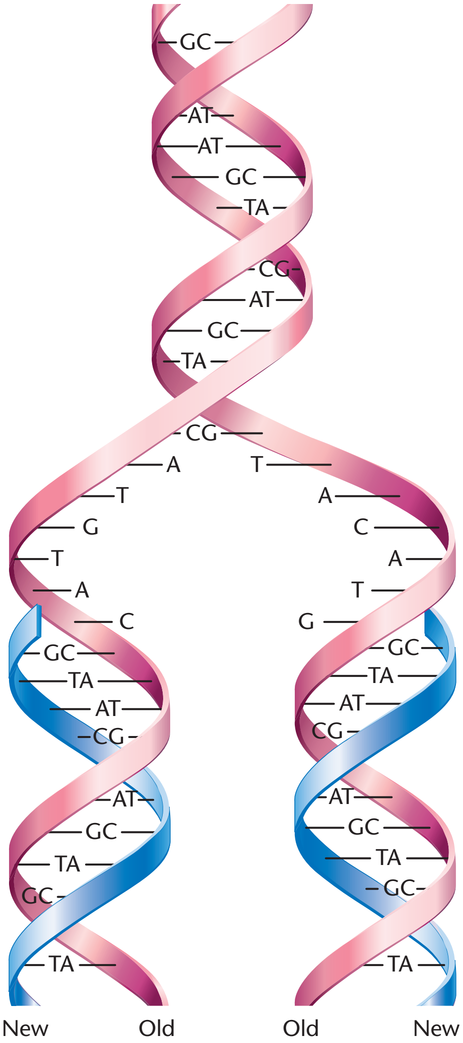


FIGURE 11.1 Generalized model of semiconservative replication of DNA. New synthesis is shown in blue.

Two other theoretical modes of replication are possible that also rely on the parental strands as a template (**Figure 11.2**). In **conservative replication,** complementary polynucleotide chains are synthesized as described earlier. Following synthesis, however, the two newly created strands then come together and the parental strands reassociate. The original helix is thus "conserved."

In the second alternative mode, called **dispersive replication,** the parental strands are dispersed into two new double helices following replication. Hence, each strand consists of both old and new DNA. This mode would involve cleavage of the parental strands during replication. It is the most complex of the three possibilities and is therefore considered to be least likely to occur. It could not, however, be ruled out as an experimental model. Figure 11.2 shows the theoretical results of a single round of replication by each of the three different modes.

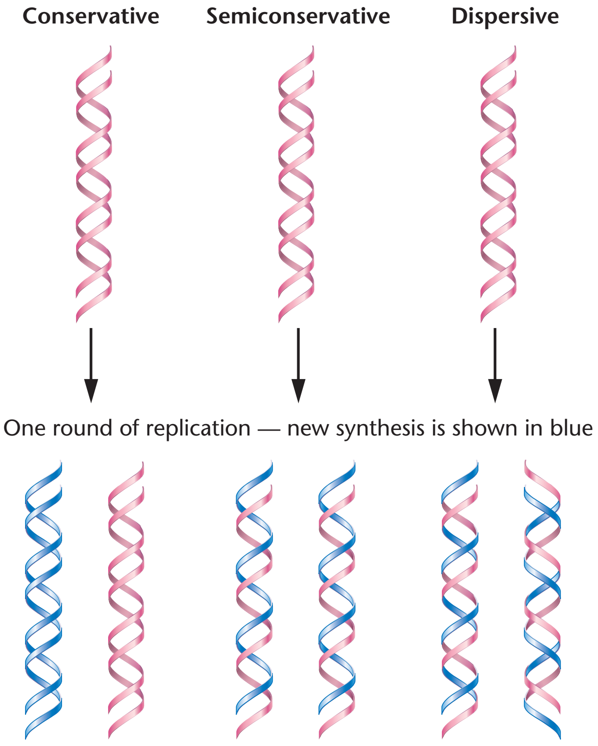


FIGURE 11.2 Results of one round of replication of DNA for each of the three possible modes by which replication could be accomplished.

The Meselson–Stahl Experiment

In 1958, Matthew Meselson and Franklin Stahl published the results of an experiment providing strong evidence that semiconservative replication is the mode used by bacterial cells to produce new DNA molecules. They grew *E. coli* cells for many generations in a medium that had $^{15}NH_4Cl$ (ammonium chloride) as the only nitrogen source. A "heavy" isotope of nitrogen, ^{15}N contains one more neutron than the naturally occurring ^{14}N isotope; thus, molecules containing ^{15}N are more dense than those containing ^{14}N. Unlike radioactive isotopes, ^{15}N is stable. After many generations in this medium, almost all nitrogen-containing molecules in the *E. coli* cells, including the nitrogenous bases of DNA, contained the heavier isotope.

Critical to the success of this experiment, DNA containing ^{15}N can be distinguished from DNA containing ^{14}N. The experimental procedure involves the use of a technique referred to as **sedimentation equilibrium centrifugation,** or as it is also called, *buoyant density gradient centrifugation.* Samples are forced by centrifugation through a density gradient of a heavy metal salt, such as cesium chloride. Molecules of DNA will reach equilibrium when their density equals the density of the gradient medium. In this case, ^{15}N-DNA will reach this point at a position closer to the bottom of the tube than will ^{14}N-DNA.

In this experiment (**Figure 11.3**), uniformly labeled ^{15}N cells were transferred to a medium containing only $^{14}NH_4Cl$. Thus, all "new" synthesis of DNA during replication contained only the "lighter" isotope of nitrogen. The time of transfer to the new medium was taken as time zero ($t = 0$). The *E. coli* cells were allowed to replicate over several generations, with cell samples removed after each replication cycle. DNA was isolated from each sample and subjected to sedimentation equilibrium centrifugation.

After one generation, the isolated DNA was present in only a single band of intermediate density—the expected result for semiconservative replication in which each replicated molecule was composed of one new ^{14}N-strand and one old ^{15}N-strand (**Figure 11.4**). This result was not consistent with the prediction of conservative replication, in which two distinct bands would occur; thus this mode may be rejected.

After two cell divisions, DNA samples showed two density bands—one intermediate band and one lighter band corresponding to the ^{14}N position in the gradient. Similar results occurred after a third generation, except that the proportion of the lighter band increased. This was again consistent with the interpretation that replication is semiconservative.

You may have realized that a molecule exhibiting intermediate density is also consistent with dispersive replication. However, Meselson and Stahl ruled out this mode of replication on the basis of two observations. First, after the first generation of replication in an ^{14}N-containing medium, they isolated the hybrid molecule and heat denatured it. Recall from earlier in the text (see Chapter 10), that heating will separate a duplex into single strands.

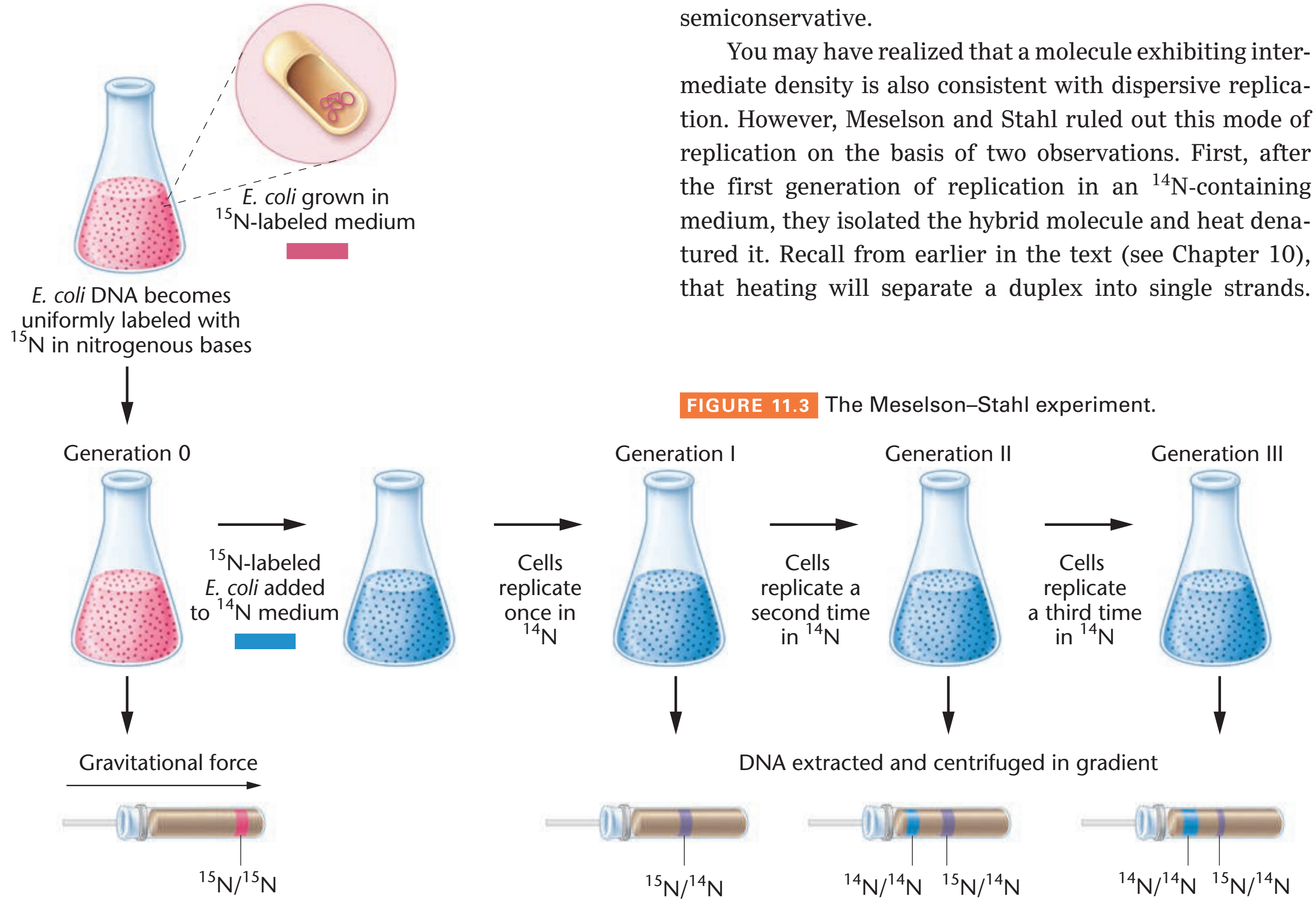

E. coli grown in ^{15}N-labeled medium
E. coli DNA becomes uniformly labeled with ^{15}N in nitrogenous bases
Generation 0
^{15}N-labeled *E. coli* added to ^{14}N medium
Cells replicate once in ^{14}N
Generation I
Cells replicate a second time in ^{14}N
Generation II
Cells replicate a third time in ^{14}N
Generation III
Gravitational force
DNA extracted and centrifuged in gradient
$^{15}N/^{15}N$
$^{15}N/^{14}N$
$^{14}N/^{14}N$
$^{15}N/^{14}N$
$^{14}N/^{14}N$
$^{15}N/^{14}N$

FIGURE 11.3 The Meselson–Stahl experiment.

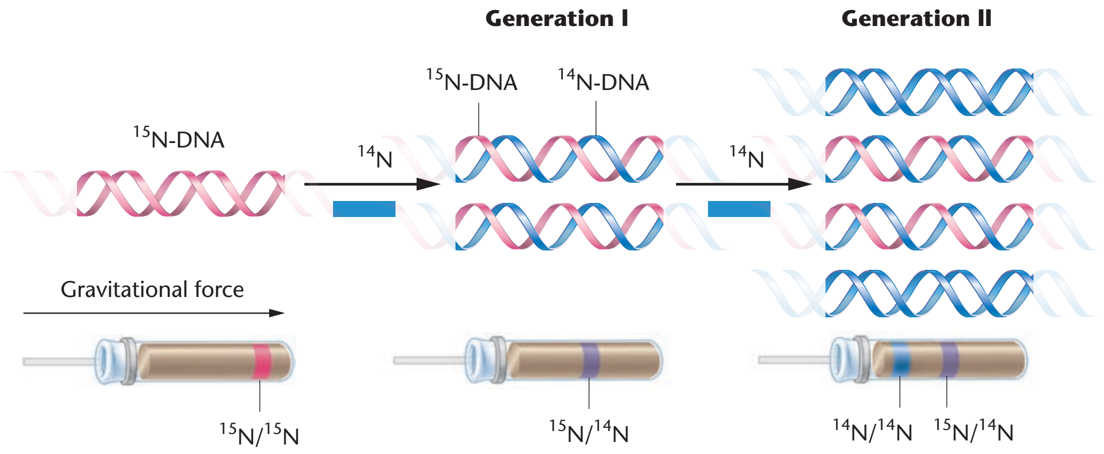


FIGURE 11.4 The expected results of two generations of semiconservative replication in the Meselson–Stahl experiment.

When the densities of the single strands of the hybrid were determined, they exhibited *either* an ^{15}N profile *or* an ^{14}N profile, but *not* an intermediate density. This observation is consistent with the semiconservative mode but inconsistent with the dispersive mode.

Furthermore, if replication were dispersive, *all* generations after $t = 0$ would demonstrate DNA of an intermediate density. In each generation after the first, the ratio of ^{15}N/^{14}N would decrease and the hybrid band would become lighter and lighter, eventually approaching the ^{14}N band. This result was not observed. The Meselson–Stahl experiment provided conclusive support for semiconservative replication in bacteria and tended to rule out both the conservative and dispersive modes.

NOW SOLVE THIS

11.1 In the Meselson–Stahl experiment, which of the three modes of replication could be ruled out after one round of replication? after two rounds?

- **HINT:** *This problem involves an understanding of the nature of the experiment as well as the difference between the three possible modes of replication. The key to its solution is to determine which mode will not create "hybrid" helices after one round of replication.*

For more practice, see Problems 3–5.

Semiconservative Replication in Eukaryotes

In 1957, the year before the work of Meselson and Stahl was published, J. Herbert Taylor, Philip Woods, and Walter Hughes presented evidence that semiconservative replication also occurs in eukaryotic organisms. They experimented with root tips of the broad bean *Vicia faba,* which are an excellent source of dividing cells. These researchers were able to monitor the process of replication by labeling DNA with ^{3}H-thymidine, a radioactive precursor of DNA, and by performing autoradiography.

Autoradiography is a common technique that, when applied cytologically, pinpoints the location of a radioisotope in a cell. In this procedure, a photographic emulsion is placed over a histological preparation containing cellular material (root tips, in this experiment), and the preparation is stored in the dark. The slide is then developed, much as photographic film is processed. Because the radioisotope emits energy, following development the emulsion turns black at the approximate point of emission. The end result is the presence of dark spots or "grains" on the surface of the section, identifying the location of newly synthesized DNA within the cell.

Taylor and his colleagues grew root tips for approximately one generation in the presence of the radioisotope and then placed them in unlabeled medium in which cell division continued. At the conclusion of each generation, they arrested the cultures at metaphase by adding colchicine (a chemical derived from the crocus plant that poisons the spindle fibers) and then examined the chromosomes by autoradiography. They found radioactive thymidine only in association with chromatids that contained newly synthesized DNA. **Figure 11.5** illustrates the replication of a single chromosome over two division cycles, including the distribution of grains.

These results are compatible with the semiconservative mode of replication. After the first replication cycle in the presence of the isotope, both sister chromatids show radioactivity, indicating that each chromatid contains one new radioactive DNA strand and one old unlabeled strand. After the second replication cycle, *which takes place in unlabeled medium,* only one of the two sister chromatids of each chromosome should be radioactive because half of the parent strands are unlabeled. With only the minor exceptions of *sister chromatid exchanges* (discussed in Chapter 5), this result was observed.

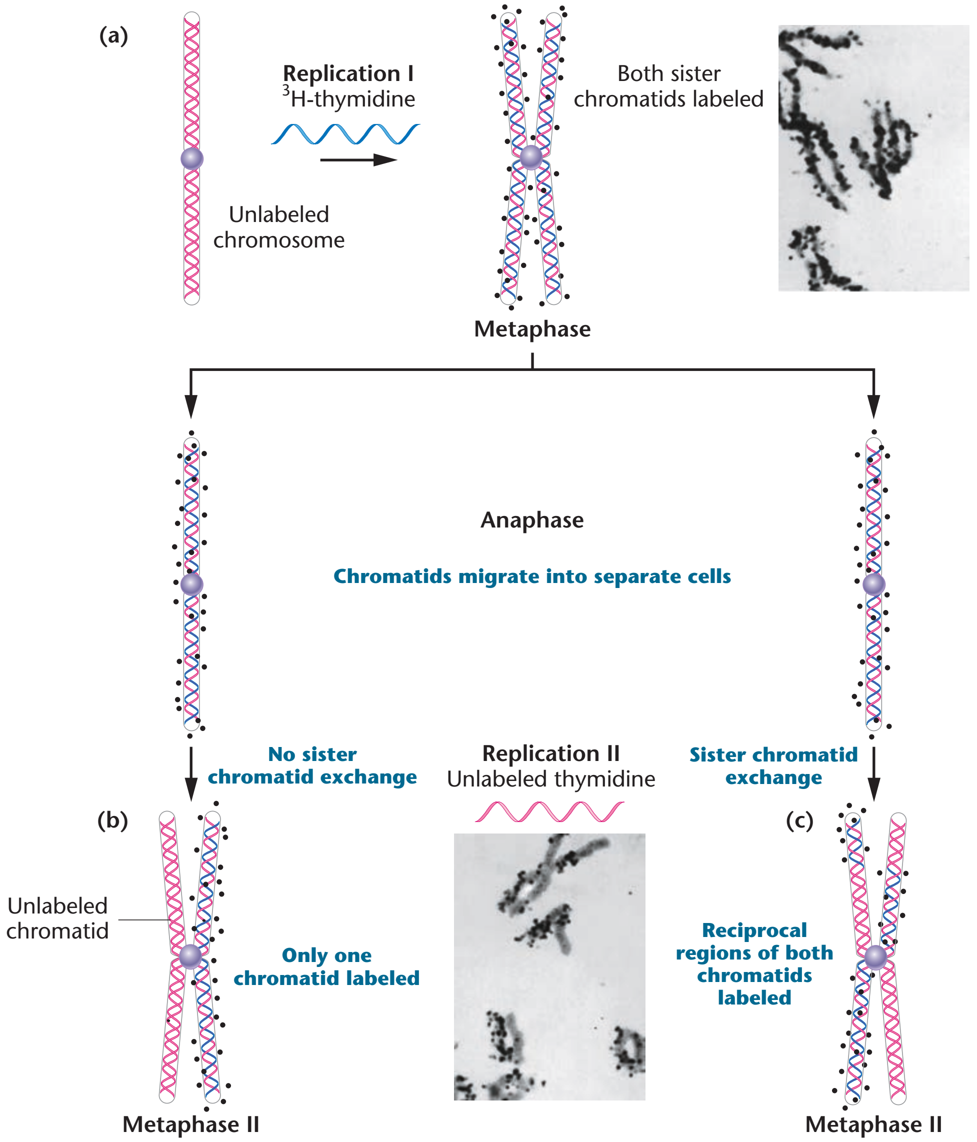

(a)
Replication I
^{3}H-thymidine
Unlabeled chromosome
Both sister chromatids labeled
Metaphase
Anaphase
Chromatids migrate into separate cells
No sister chromatid exchange
Replication II
Unlabeled thymidine
Sister chromatid exchange
(b)
Unlabeled chromatid
Only one chromatid labeled
(c)
Reciprocal regions of both chromatids labeled
Metaphase II
Metaphase II

FIGURE 11.5 The Taylor–Woods–Hughes experiment, demonstrating the semiconservative mode of replication of DNA in root tips of *Vicia faba*. (a) An unlabeled chromosome proceeds through the cell cycle in the presence of ^{3}H-thymidine. As it enters mitosis, both sister chromatids of the chromosome are labeled, as shown, by autoradiography. After a second round of replication (b), this time in the absence of ^{3}H-thymidine, only one chromatid of each chromosome is expected to be surrounded by grains. Except where a reciprocal exchange has occurred between sister chromatids (c), the expectation was upheld. The micrographs are of the actual autoradiograms obtained in the experiment.

Together, the Meselson–Stahl experiment and the experiment by Taylor, Woods, and Hughes soon led to the general acceptance of the semiconservative mode of replication. Later studies with other organisms reached the same conclusion and also strongly supported Watson and Crick's proposal for the double-helix model of DNA.

Origins, Forks, and Units of Replication

To enhance our understanding of semiconservative replication, let's briefly consider a number of relevant issues. The first concerns the **origin of replication.** Where along the chromosome is DNA replication initiated? Is there only a single origin, or does DNA synthesis begin at more than one point? Is any given point of origin random, or is it located at a specific region along the chromosome? Second, once replication begins, does it proceed in a single direction or in both directions away from the origin? In other words, is replication *unidirectional* or *bidirectional*?

To address these issues, we need to introduce two terms. First, at each point along the chromosome where replication is occurring, the strands of the helix are unwound, creating what is called a **replication fork.** Such a fork will initially appear at the point of origin of synthesis and then move along the DNA duplex as replication proceeds. If replication is bidirectional, two such forks will be present, migrating in opposite directions away from the origin. The second term refers to the length of DNA that is replicated following one initiation event at a single origin. This is a unit referred to as the **replicon.**

The evidence is clear regarding the origin and direction of replication. John Cairns tracked replication in *E. coli,* using radioactive precursors of DNA synthesis and autoradiography. He was able to demonstrate that in *E. coli* there is only a single region, called *oriC,* where replication is initiated. The presence of only a single origin is characteristic of bacteria, which have only one circular chromosome. Since DNA synthesis in bacteria originates at a single point, the entire chromosome constitutes one replicon. In *E. coli,* the replicon consists of the entire genome of 4.6 Mb (4.6 million base pairs).

Figure 11.6 illustrates Cairns's interpretation of DNA replication in *E. coli.* This interpretation does not answer the question of unidirectional versus bidirectional synthesis. However, other results, derived from studies of bacteriophage lambda, demonstrated that replication is bidirectional, moving away from *oriC* in both directions. Figure 11.6 therefore interprets Cairns's work with that understanding. Bidirectional replication creates two replication forks that migrate farther and farther apart as replication proceeds. These forks eventually merge, as semiconservative replication of the entire chromosome is completed, at a termination region, called *ter.*

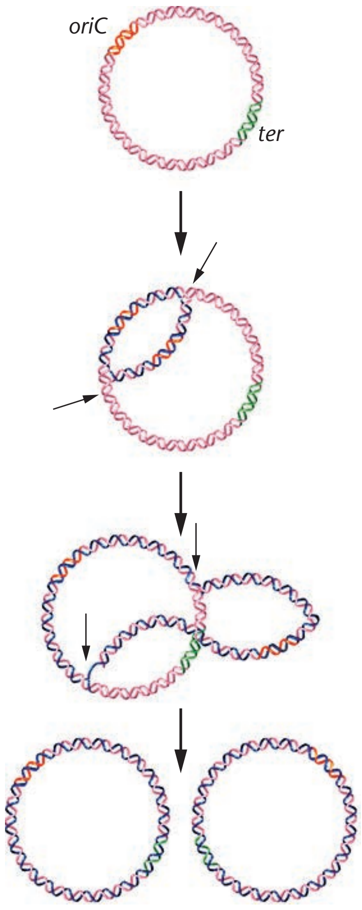


FIGURE 11.6 Bidirectional replication of the *E. coli* chromosome. The thin black arrows identify the advancing replication forks.

Later in this chapter we will see that in eukaryotes, each chromosome contains multiple points of origin.

11.2 DNA Synthesis in Bacteria Involves Five Polymerases, as Well as Other Enzymes

To say that replication is semiconservative and bidirectional describes the overall *pattern* of DNA duplication and the association of finished strands with one another once synthesis is completed. However, it says little about the more complex issue of how the actual *synthesis* of long complementary polynucleotide chains occurs on a DNA template. Like most questions in molecular biology, this one was first studied using microorganisms. Research on DNA synthesis began about the same time as the Meselson–Stahl work, and the topic is still an active area of investigation. What is most apparent in this research is the tremendous complexity of the biological synthesis of DNA.

DNA Polymerase I

Studies of the enzymology of DNA replication were first reported by Arthur Kornberg and colleagues in 1957. They isolated an enzyme from *E. coli* that was able to direct DNA synthesis in a cell-free (*in vitro*) system. The enzyme is called **DNA polymerase I,** because it was the first of several similar enzymes to be isolated.

Kornberg determined that there were two major requirements for *in vitro* DNA synthesis under the direction of DNA polymerase I: (1) all four deoxyribonucleoside triphosphates (dNTPs) and (2) template DNA. If any one of the four deoxyribonucleoside triphosphates was omitted from the reaction, no measurable synthesis occurred. If derivatives of these precursor molecules other than the nucleoside triphosphate were used (nucleotides or nucleoside diphosphates), synthesis also did not occur. If no template DNA was added, synthesis of DNA occurred but was reduced greatly.

Most of the synthesis directed by Kornberg's enzyme appeared to be exactly the type required for semiconservative replication. The reaction is summarized in **Figure 11.7**, which depicts the addition of a single nucleotide. The enzyme has since been shown to consist of a single polypeptide containing 928 amino acids.

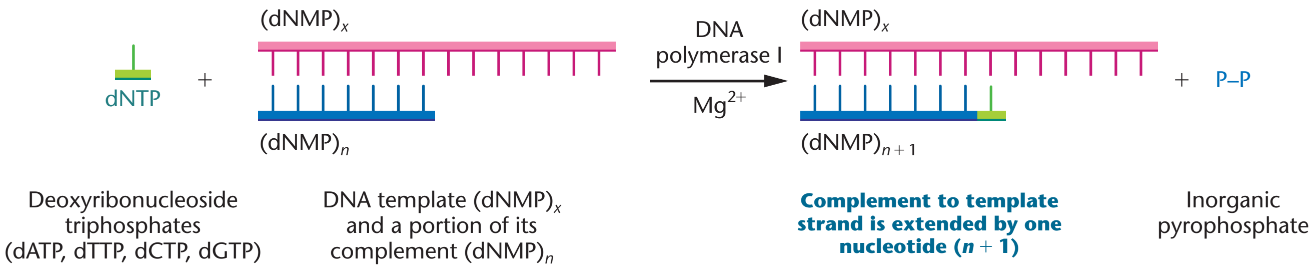

(dNMP)$_x$
dNTP
+
(dNMP)$_n$
DNA polymerase I
Mg^{2+}
(dNMP)$_x$
(dNMP)$_{n+1}$
+
P–P
Deoxyribonucleoside triphosphates (dATP, dTTP, dCTP, dGTP)
DNA template (dNMP)$_x$ and a portion of its complement (dNMP)$_n$
Complement to template strand is extended by one nucleotide (n + 1)
Inorganic pyrophosphate

FIGURE 11.7 The chemical reaction catalyzed by DNA polymerase I. During each step, a single nucleotide is added to the growing complement of the DNA template, using a nucleoside triphosphate as the substrate. The release of inorganic pyrophosphate drives the reaction energetically.

The way in which each nucleotide is added to the growing chain is a function of the specificity of DNA polymerase I. As shown in Figure 11.8, the precursor dNTP contains the three phosphate groups attached to the 5′ carbon of deoxyribose. As the two terminal phosphates are cleaved during synthesis, the remaining phosphate attached to the 5′ carbon is covalently linked to the 3′-OH group of the deoxyribose to which it is added. Thus, **chain elongation** occurs in the **5′ to 3′ direction** by the addition of one nucleotide at a time to the growing 3′ end. Each step provides a newly exposed 3′-OH group that can participate in the next addition of a nucleotide as DNA synthesis proceeds.

Having isolated DNA polymerase I and established its catalytic activity, Kornberg next sought to demonstrate the accuracy, or fidelity, with which the enzyme replicated the DNA template. Because technology for ascertaining the nucleotide sequences of the template and newly synthesized strand was not yet available in 1957, he initially had to rely on several indirect methods.

One of Kornberg's approaches was to compare the nitrogenous base compositions of the DNA template with those of the recovered DNA product. Table 11.1 shows Kornberg's base-composition analysis of three DNA templates. Within experimental error, the base composition of each product agreed with the template DNAs used. These data, along with other types of comparisons of template and product, suggested that the templates were replicated faithfully.

DNA Polymerase II, III, IV, and V

While DNA polymerase I (DNA polymerases are abbreviated as DNA Pol I, II, etc.) clearly directs the synthesis of DNA, a serious reservation about the enzyme's true biological role was raised in 1969. Paula DeLucia and John Cairns discovered a mutant strain of *E. coli* that was deficient in polymerase I activity. The mutation was designated *polA1*. In the absence of the functional enzyme, this mutant strain of *E. coli* still duplicated its DNA and successfully reproduced. However, the cells were deficient in their ability to

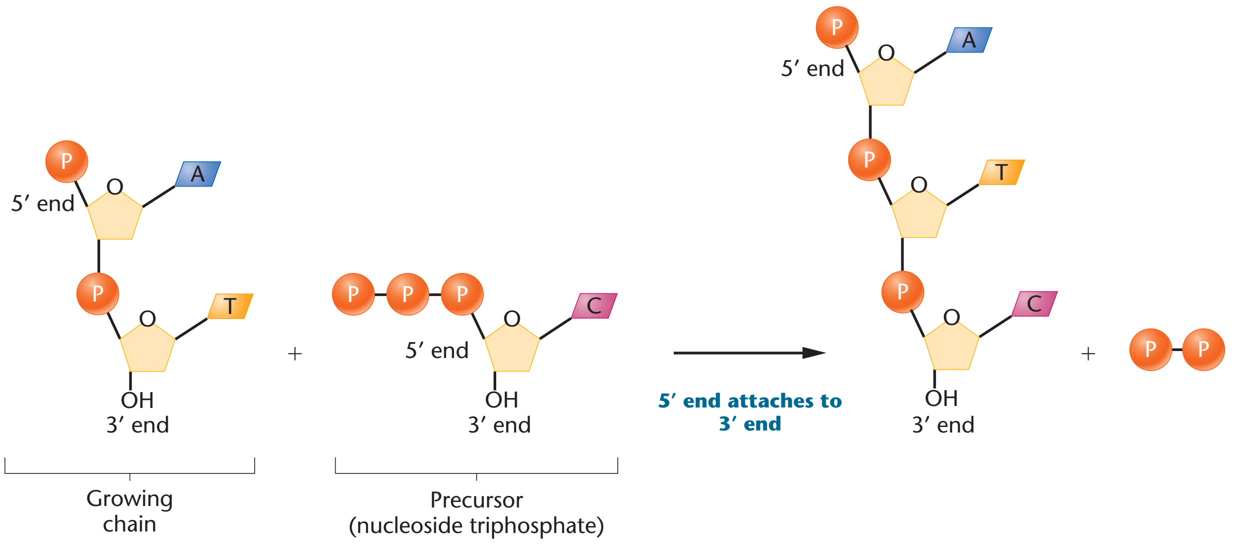


FIGURE 11.8 Demonstration of 5′ to 3′ synthesis of DNA.

TABLE 11.1 Base Composition of the DNA Template and the Product of Replication in Kornberg's Early Work

Organism	Template or Product	%A	%T	%G	%C
T2	Template	32.7	33.0	16.8	17.5
	Product	33.2	32.1	17.2	17.5
E. coli	Template	25.0	24.3	24.5	26.2
	Product	26.1	25.1	24.3	24.5
Calf	Template	28.9	26.7	22.8	21.6
	Product	28.7	27.7	21.8	21.8

Source: Kornberg (1960).

repair DNA. For example, the mutant strain is highly sensitive to ultraviolet (UV) light and radiation, both of which damage DNA and are mutagenic. Nonmutant bacteria are able to repair a great deal of UV-induced damage.

These observations led to two conclusions:

1. At least one enzyme other than DNA Pol I is responsible for replicating DNA *in vivo* in *E. coli* cells.
2. DNA Pol I serves a secondary function *in vivo* and is now believed to be critical to the maintenance of fidelity of DNA synthesis.

To date, four other unique DNA polymerases have been isolated from cells lacking polymerase I activity and from normal cells that contain DNA Pol I. **Table 11.2** contrasts several characteristics of DNA Pol I with **DNA Pol II** and **III.** Although none of the three can *initiate* DNA synthesis on a template, all three can *elongate* an existing DNA strand, called a **primer,** and all possess 3′ to 5′ exonuclease activity, which means that they have the potential to polymerize in one direction and then pause, reverse their direction, and excise nucleotides just added. As we will later discuss, this activity provides a capacity to proofread newly synthesized DNA and to remove and replace incorrect nucleotides.

DNA Pol I also demonstrates 5′ to 3′ exonuclease activity. This activity allows the enzyme to excise nucleotides, starting at the end at which synthesis begins and proceeding in the same direction of synthesis. Two final observations probably explain why Kornberg isolated polymerase I and not polymerase III: polymerase I is present in greater amounts than is polymerase III, and it is also much more stable.

What then are the roles of the polymerases *in vivo*? DNA Pol III is the enzyme responsible for the 5′ to 3′ polymerization essential to *in vivo* replication. Its 3′ to 5′ exonuclease activity also provides a proofreading function that is activated when it inserts an incorrect nucleotide. When this occurs, synthesis stalls and the polymerase "reverses course," excising the incorrect nucleotide. Then, it proceeds back in the 5′ to 3′ direction, synthesizing the complement of the template strand. DNA Pol I is believed to be responsible for removing the primer, as well as for the synthesis that fills gaps produced after this removal. Its exonuclease activities also allow for its participation in DNA repair. DNA Pol II, as well as **DNA Pol IV** and **V,** are involved in various aspects of repair of DNA that has been damaged by external forces, such as ultraviolet light. Polymerase II is encoded by a gene activated by disruption of DNA synthesis at the replication fork.

TABLE 11.2 Properties of Bacterial DNA Polymerases I, II, and III

Properties	I	II	III
Initiation of chain synthesis	–	–	–
5′–3′ polymerization	+	+	+
3′–5′ exonuclease activity	+	+	+
5′–3′ exonuclease activity	+	–	–
Molecules of polymerase/cell	400	?	15

The DNA Pol III Holoenzyme

We conclude this section by emphasizing the complexity of the DNA Pol III molecule. The active form of DNA Pol III, referred to as the **holoenzyme,** is made up of unique polypeptide subunits, ten of which have been identified (**Table 11.3**). The largest subunit, α, along with subunits ϵ and θ, form a complex called the **core enzyme,** which imparts the catalytic function to the holoenzyme. In *E. coli*, each holoenzyme contains two, and possibly three, core enzyme complexes. As part of each core, the α subunit is responsible for DNA synthesis along the template strands, whereas the ϵ subunit possesses 3′ to 5′ exonuclease capability, essential to proofreading. The need for more than one core enzyme will soon become apparent. A second group of five subunits (γ, δ, δ', χ, and ν) are complexed to form what is called the **sliding clamp loader,** which pairs with the core enzyme and facilitates the function of a critical component of the holoenzyme, called the **sliding DNA clamp.** The enzymatic function of the sliding clamp loader is dependent on energy generated by the hydrolysis of ATP. The sliding DNA clamp links to the core enzyme and is made up of multiple copies of the β subunit, taking on the

TABLE 11.3 Subunits of the DNA Polymerase III Holoenzyme

Subunit	Function	Groupings
α	5′–3′ polymerization	Core enzyme: Elongates polynucleotide chain and proofreads
ϵ	3′–5′ exonuclease	
θ	Core assembly	
γ δ δ′ χ ν	Loads enzyme on template (serves as clamp loader)	γ complex
β	Sliding clamp structure (processivity factor)	
τ	Dimerizes core complex	

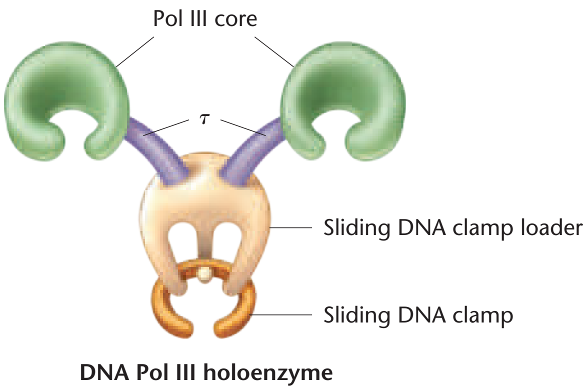


FIGURE 11.9 The components making up the DNA Pol III holoenzyme, as described in the text. While there may be three core enzyme complexes present in the holoenzyme, for simplicity, we illustrate only two.

shape of a donut, whereby it can open and shut, to encircle the unreplicated DNA helix. By doing so, and being linked to the core enzyme, the clamp leads the way during synthesis, maintaining the binding of the core enzyme to the template during polymerization of nucleotides. Thus, the length of DNA that is replicated by the core enzyme before it detaches from the template, a property referred to as **processivity,** is vastly increased. There is one sliding clamp per core enzyme. Finally, one τ subunit interacts with each core enzyme, linking it to the sliding clamp loader.

The DNA Pol III holoenzyme is diagrammatically illustrated in Figure 11.9. You should compare the diagram to the description of each component above. Note that we have shown the holoenzyme to contain two core enzyme complexes, although as stated above, a third one may be present. The components of the DNA Pol III holoenzyme will be referred to in the discussion that follows.

11.3 Many Complex Issues Must Be Resolved during DNA Replication

We have thus far established that in bacteria and viruses replication is semiconservative and bidirectional along a single replicon. We also know that synthesis is catalyzed by DNA polymerase III and occurs in the 5′ to 3′ direction. Bidirectional synthesis creates two replication forks that move in opposite directions away from the origin of synthesis. As we can see from the following list, many issues remain to be resolved in order to provide a comprehensive understanding of DNA replication:

1. The helix must undergo localized unwinding, and the resulting “open” configuration must be stabilized so that synthesis may proceed along both strands.
2. As unwinding and subsequent DNA synthesis proceed, increased coiling creates tension further down the helix, which must be reduced.
3. A primer of some sort must be synthesized so that polymerization can commence under the direction of DNA polymerase III. Surprisingly, RNA, not DNA, serves as the primer.
4. Once the RNA primers have been synthesized, DNA polymerase III begins to synthesize the DNA complement of both strands of the parent molecule. Because the two strands are antiparallel to one another, continuous synthesis in the direction that the replication fork moves is possible along only one of the two strands. On the other strand, synthesis must be discontinuous and thus involves a somewhat different process.
5. The RNA primers must be removed prior to completion of replication. The gaps that are temporarily created must be filled with DNA complementary to the template at each location.
6. The newly synthesized DNA strand that fills each temporary gap must be joined to the adjacent strand of DNA.
7. While DNA polymerases accurately insert complementary bases during replication, they are not perfect, and, occasionally, incorrect nucleotides are added to the growing strand. A proofreading mechanism that also corrects errors is an integral process during DNA synthesis.

As we consider these points, examine Figures 11.10, 11.11, and 11.12 to see how each issue is resolved. Figure 11.13 summarizes the model of DNA synthesis.

Unwinding the DNA Helix

As discussed earlier, there is a single point of origin along the circular chromosome of most bacteria and viruses at which DNA replication is initiated. This region in *E. coli* has been particularly well studied and is called ***oriC.*** It consists of 245 DNA base pairs and is characterized by five repeating sequences of 9 base pairs, and three repeating sequences of 13 base pairs, called **9mers** and **13mers,** respectively. Both 9mers and 13mers are AT-rich, which renders them relatively less stable than an average double-helical sequence of DNA, which no doubt enhances helical unwinding. A specific initiator protein, called **DnaA** (because it is encoded by the *dnaA* gene), is responsible for initiating replication by binding to a region of 9mers. This newly formed complex then undergoes a slight conformational change and associates with the region of 13mers, which causes the helix to destabilize and open up, exposing single-stranded regions of DNA (ssDNA). This step facilitates the subsequent binding of another key player in the process—a protein called

DNA helicase (made up of multiple copies of the DnaB polypeptide). DNA helicase is assembled as a hexamer of subunits around one of the exposed single-stranded DNA molecules. The helicase subsequently recruits the holoenzyme to bind to the newly formed replication fork to formally initiate replication, and it then proceeds to move along the ssDNA, opening up the helix as it progresses. Helicases require energy supplied by the hydrolysis of ATP, which aids in denaturing the hydrogen bonds that stabilize double helix.

Once the helicase has opened up the helix and ssDNA is available, base pairing must be inhibited until it can serve as a template for synthesis. This is accomplished by proteins that bind specifically to single strands of DNA, appropriately called **single-stranded binding proteins** (**SSBs**).

As unwinding proceeds, a coiling tension is created ahead of the replication fork, often producing **supercoiling.** In circular molecules, supercoiling may take the form of added twists and turns of the DNA, much like the coiling you can create in a rubber band by stretching it out and then twisting one end. Such supercoiling can be relaxed by **DNA gyrase,** a member of a larger group of enzymes referred to as **DNA topoisomerases.** The gyrase makes either single- or double-stranded "cuts" and also catalyzes localized movements that have the effect of "undoing" the twists and knots created during supercoiling. The strands are then resealed. These various reactions are driven by the energy released during ATP hydrolysis.

Together, the DNA, the polymerase complex, and associated enzymes make up an array of molecules that initiate DNA synthesis and are part of what we have previously called the *replisome*.

Initiation of DNA Synthesis Using an RNA Primer

Once a small portion of the helix is unwound and the holoenzyme is in place, how is synthesis initiated? As we have seen, DNA polymerase III requires a primer with a free 3′-hydroxyl group in order to elongate a polynucleotide chain. Since none is available in a circular chromosome, this absence prompted researchers to investigate how the first nucleotide could be added. It is now clear that RNA serves as the primer that initiates DNA synthesis.

A short segment of RNA (about 10 to 12 nucleotides long), complementary to DNA, is first synthesized on the DNA template. Synthesis of the RNA is directed by a form of RNA polymerase called **primase,** which is recruited to the replication fork by DNA helicase, and which does not require a free 3′ end to initiate synthesis. It is to this short segment of RNA that DNA Pol III begins to add deoxyribonucleotides, initiating DNA synthesis. A conceptual diagram of initiation on a DNA template is shown in Figure 11.10. Later, the RNA primer is clipped out and replaced with DNA. This is thought to occur under the direction of DNA Pol I. Recognized in viruses, bacteria, and several eukaryotic organisms, RNA priming is a universal phenomenon during the initiation of DNA synthesis.

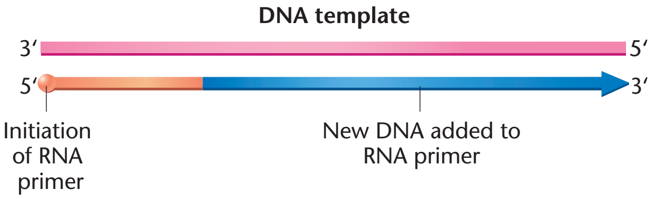


FIGURE 11.10 The initiation of DNA synthesis. A complementary RNA primer is first synthesized, to which DNA is added. All synthesis is in the 5′ to 3′ direction. Eventually, the RNA primer is replaced with DNA under the direction of DNA polymerase I.

Continuous and Discontinuous DNA Synthesis

We must now revisit the fact that the two strands of a double helix are **antiparallel** to each other—that is, one runs in the 5′ to 3′ direction, while the other has the opposite 3′ to 5′ polarity. Because DNA Pol III synthesizes DNA in only the 5′ to 3′ direction, synthesis along an advancing replication fork occurs in one direction on one strand and in the opposite direction on the other.

As a result, as the strands unwind and the replication fork progresses down the helix (Figure 11.11), only one strand can serve as a template for **continuous DNA synthesis.** This newly synthesized DNA is called the **leading strand.** As the fork progresses, many points of initiation are necessary on the opposite DNA template, resulting in **discontinuous DNA synthesis** of the **lagging strand.***

Evidence supporting the occurrence of discontinuous DNA synthesis was first provided by Reiji and Tuneko Okazaki. They discovered that when bacteriophage DNA is replicated in *E. coli,* some of the newly formed DNA that is hydrogen bonded to the template strand is present as small fragments containing 1000 to 2000 nucleotides. RNA primers are part of each such fragment. These pieces, now called **Okazaki fragments,** are converted into longer and longer DNA strands of higher molecular weight as synthesis proceeds.

Discontinuous synthesis of DNA requires enzymes that both remove the RNA primers and unite the Okazaki fragments into the lagging strand. As we have noted, DNA Pol I removes the primers and replaces the missing nucleotides. Joining the fragments is the work of another enzyme, **DNA ligase,** which is capable of catalyzing the

*Because DNA synthesis is continuous on one strand and discontinuous on the other, the term **semidiscontinuous synthesis** is sometimes used to describe the overall process.

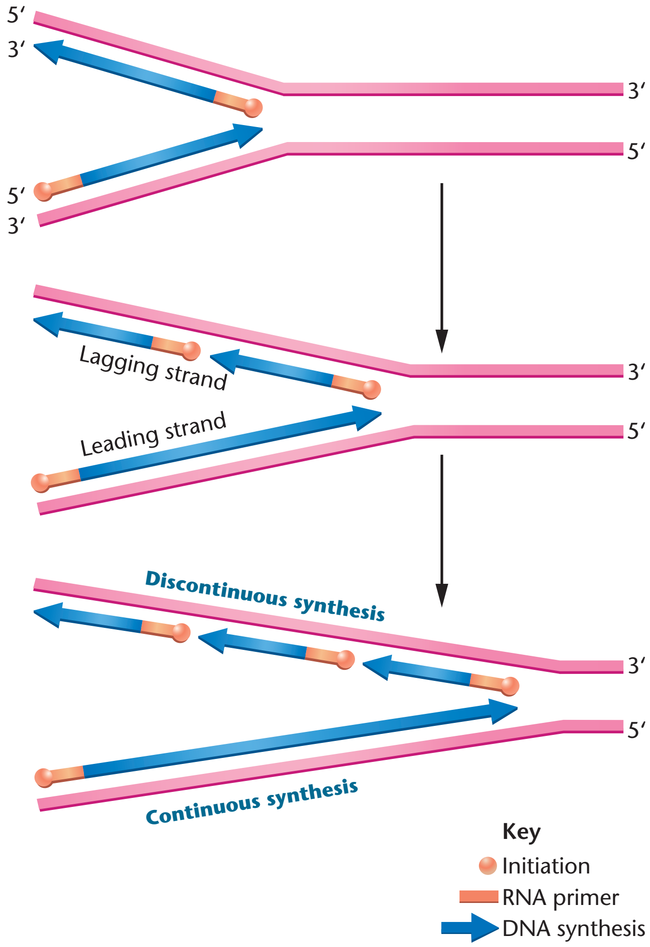


FIGURE 11.11 Opposite polarity of synthesis along the two strands of DNA is necessary because they run antiparallel to one another, and because DNA polymerase III synthesizes in only one direction (5′ to 3′). On the lagging strand, synthesis must be discontinuous, resulting in the production of Okazaki fragments. On the leading strand, synthesis is continuous. RNA primers are used to initiate synthesis on both strands.

formation of the phosphodiester bond that seals the nick between the discontinuously synthesized strands. The evidence that DNA ligase performs this function during DNA synthesis is strengthened by the observation of a ligase-deficient mutant strain (*lig*) of *E. coli,* in which a large number of unjoined Okazaki fragments accumulate.

Concurrent Synthesis Occurs on the Leading and Lagging Strands

Given the model just discussed, we might ask how the holoenzyme of DNA Pol III synthesizes DNA on both the leading and lagging strands. Can both strands be replicated simultaneously at the same replication fork, or are the events distinct, involving two separate copies of the enzyme? Evidence suggests that both strands are replicated simultaneously, with each strand acted upon by one of the two core enzymes that are part of the DNA Pol III holoenzyme. As **Figure 11.12** illustrates, if the lagging strand is spooled out, forming a loop, nucleotide polymerization can occur simultaneously on both template strands under the direction of the holoenzyme. After the synthesis of 1000 to 2000 nucleotides, the monomer of the enzyme on the lagging strand will encounter a completed Okazaki fragment, at which point it releases the lagging strand. A new loop of the lagging strand is spooled out, and the process is repeated. Looping inverts the orientation of the template but not the direction of actual synthesis on the lagging strand, which is always in the 5′ to 3′ direction. As mentioned earlier, it is believed that there is a third core enzyme associated with the DNA Pol III holoenzyme, and that it functions in the synthesis of Okazaki fragments. For simplicity, we will include only two core enzymes in this and subsequent figures.

Another important feature of the holoenzyme that facilitates synthesis at the replication fork is the donut-shaped sliding DNA clamp that surrounds the unreplicated double helix and is linked to the advancing core enzyme. This clamp prevents the core enzyme from dissociating from the template as polymerization proceeds. By doing so, the clamp is responsible for vastly increasing the processivity of the core enzyme—that is, the number of nucleotides that may be continually added prior to dissociation from the template. This function is critical to the rapid *in vivo* rate of DNA synthesis during replication.

Proofreading and Error Correction Occurs during DNA Replication

The immediate purpose of DNA replication is the synthesis of a new strand that is precisely complementary to the template strand at each nucleotide position. Although the action of DNA polymerases is very accurate, synthesis is not perfect and a noncomplementary nucleotide is occasionally inserted erroneously. To compensate for such inaccuracies, the DNA polymerases all possess 3′ to 5′ exonuclease activity. This property imparts the potential for them to detect and excise a mismatched nucleotide (in the 3′ to 5′ direction). Once the mismatched nucleotide is removed, 5′ to 3′ synthesis can again proceed. This process, called **proofreading,** increases the fidelity of synthesis by a factor of about 100. In the case of the holoenzyme form of DNA polymerase III, the epsilon (ϵ) subunit of the core enzyme is directly involved in the proofreading step. In strains of *E. coli* with a mutation that has rendered the ϵ subunit nonfunctional, the error rate (the mutation rate) during DNA synthesis is increased substantially.

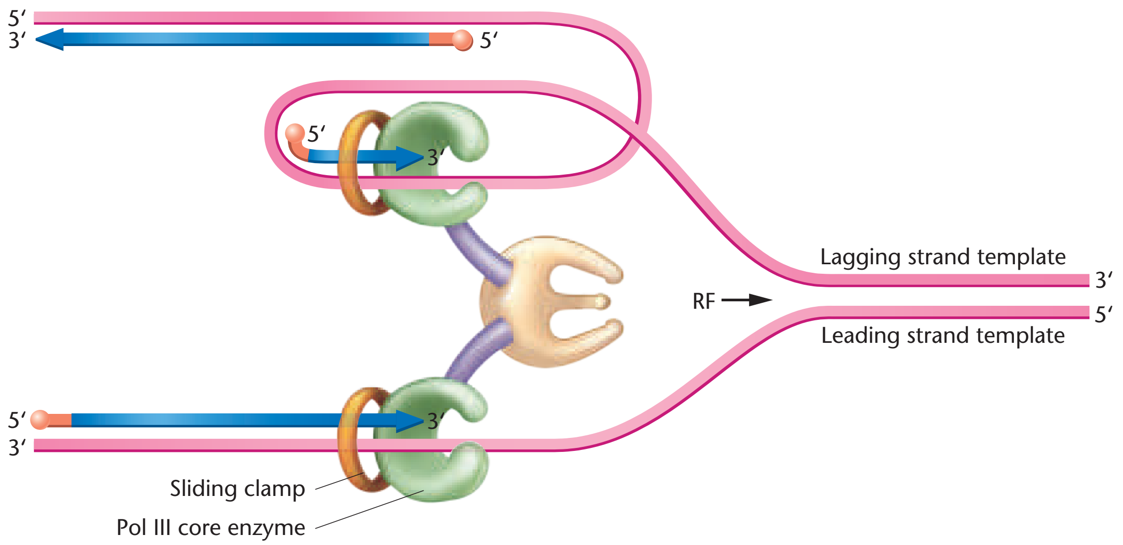


FIGURE 11.12 Illustration of how concurrent DNA synthesis may be achieved on both the leading and lagging strands at a single replication fork (RF). The lagging template strand is "looped" in order to invert the physical direction of synthesis, but not the biochemical direction. The enzyme functions as a dimer, with each core enzyme achieving synthesis on one or the other strand.

11.4 A Coherent Model Summarizes DNA Replication

We can now combine the various aspects of DNA replication occurring at a single replication fork into a coherent model, as shown in Figure 11.13. At the advancing fork, a helicase is unwinding the double helix. Once unwound, single-stranded binding proteins associate with the strands, preventing the reformation of the helix. In advance of the replication fork, DNA gyrase functions to diminish the tension created as the helix supercoils. Each of the core enzymes of DNA Pol III holoenzyme is bound to one of the template strands by a sliding DNA clamp. Continuous synthesis occurs on the leading strand, while the lagging strand must loop around in order for simultaneous (concurrent) synthesis to occur on both strands. Not shown in the figure, but essential to replication on the lagging strand, is the action of DNA polymerase I and DNA ligase, which together replace the RNA primers with DNA and join the Okazaki fragments, respectively.

The above model provides a summary of DNA synthesis against which genetic phenomena can be interpreted.

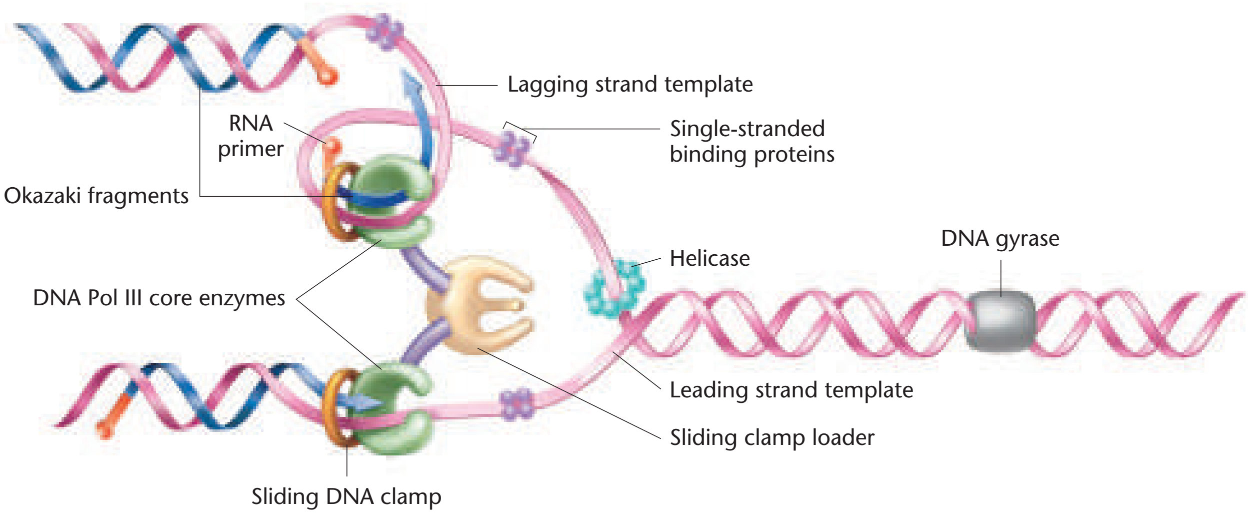


FIGURE 11.13 Summary of DNA synthesis at a single replication fork. Various enzymes and proteins essential to the process are shown.

NOW SOLVE THIS

11.2 An alien organism was investigated. When DNA replication was studied, a unique feature was apparent: No Okazaki fragments were observed. Create a model of DNA that is consistent with this observation.

- **HINT:** *This problem involves an understanding of the process of DNA synthesis in bacteria, as depicted in Figure 11.13. The key to its solution is to consider why Okazaki fragments are observed during DNA synthesis and how their formation relates to DNA structure, as described in the Watson–Crick model.*

11.5 Replication Is Controlled by a Variety of Genes

Much of what we know about DNA replication in viruses and bacteria is based on genetic analysis of the process. For example, we have already discussed studies involving the *polA1* mutation, which revealed that DNA polymerase I is not the major enzyme responsible for replication. Many other mutations interrupt or seriously impair some aspect of replication, such as the ligase-deficient and the proofreading-deficient mutations mentioned previously. Because such mutations are lethal ones, genetic analysis frequently uses **conditional mutations,** which are expressed under one condition but not under a different condition. For example, a **temperature-sensitive mutation** may not be expressed at a particular *permissive* temperature. When mutant cells are grown at a *restrictive* temperature, the mutant phenotype is expressed and can be studied. By examining the effect of the loss of function associated with the mutation, the investigation of such temperature-sensitive mutants can provide insight into the product and the associated function of the normal, nonmutant gene.

As shown in **Table 11.4**, a variety of genes in *E. coli* specify the subunits of the DNA polymerases and encode products involved in specification of the origin of synthesis, helix unwinding and stabilization, initiation and priming, relaxation of supercoiling, repair, and ligation. The discovery of such a large group of genes attests to the complexity of the process of replication, even in this relatively simple organism. Given the enormous quantity of DNA that must be unerringly replicated in a very brief time, this level of complexity is not unexpected. As we will see in Section 11.6, the process is even more involved and therefore more difficult to investigate in eukaryotes.

TABLE 11.4 Some of the Various *E. coli* Genes and Their Products or Role in Replication

Gene	Product or Role
polA	DNA polymerase I
polB	DNA polymerase II
dnaE, N, Q, X, Z	DNA polymerase III subunits
dnaG	Primase
dnaA, I, P	Initiation
dnaB, C	Helicase at *oriC*
gyrA, B	Gyrase subunits
lig	DNA ligase
rep	DNA helicase
ssb	Single-stranded binding proteins
rpoB	RNA polymerase subunit

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Lethal Knockouts

In this chapter you learned about the essential role of DNA ligase during DNA replication. Ligases also play important roles in DNA recombination and repair. Eukaryotic cells express at least three different DNA ligase genes (*Lig1*, *Lig3,* and *Lig4*) with related sequences and structures, but each enzyme has different functions. DNA ligase 1 (*Lig1*) is the key ligase involved in DNA replication in mammalian cells, *Lig3* functions in the repair of DNA single-stranded breaks and the excision repair process (discussed in Chapter 15), and *Lig4* is necessary for recombination and a process called nonhomologous end-joining. How do we know these DNA ligases exist, and how do we know their specific roles? Here we illustrate an experimental approach called **gene knockout** that reveals how organisms are affected by the designed loss of specific genes.

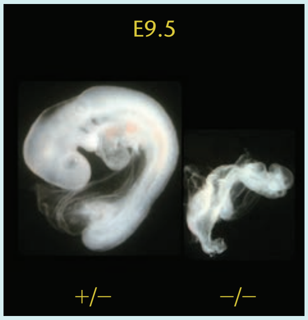
To understand the roles of DNA ligases, scientists have generated ligase knockout mouse strains that have each lost a specific ligase gene. Knockout mice are made by genetically engineering parents that are heterozygous for a particular gene; one allele is wild type, and the other is a "null," or nonfunctional, allele. These animals are denoted as +/−. When the two heterozygotes are bred together, some homozygous −/− (null) offspring are created and selected for study. When −/− offspring are not produced from a mating, scientists often suspect problems with embryonic lethality. It is often a lethality in knockouts that reveals the functional significance of a gene during embryonic development.

Here we briefly highlight examples of results from ligase knockout experiments.

Results:

Experiments from several different groups have shown that:

- Embryos from *Lig1* knockout mice develop normally to midterm before dying.
- *Lig3* knockout mice terminate embryonic development at 8.5 days postcoitum (dpc). Excessive cell death (apoptosis) occurs at 9.5 dpc. Thus *Lig3* null mice also show embryonic lethality. This lethality has been traced to an essential role of *Lig3* in mitochondria.
- *Lig4* knockout mice also show embryonic lethality. *Lig4*−/− cultured cell lines show increased sensitivity to damage by ionizing radiation and are deficient in nonhomologous end-joining repair of double-stranded DNA breaks.



Lig3 *deficiency results in excessive apoptosis at day 9.5. (Left) Heterozygous* (+/−) *mouse embryos show wild-type embryonic development whereas* Lig 3 −/− *null mouse embryos (right) arrest at 9.5 dpc.*

Conclusions:

All three of these mouse knockouts resulted in embryonic lethality. How do these findings help us understand what each of the ligases do? One conclusion we can draw right away is that each ligase is essential and plays a role independent of that of the other ligases, based on the observation that the other ligases cannot substitute for it. Thus, mutation or loss of any DNA ligase gene cannot be fully compensated for by another DNA ligase gene.

These experiments also demonstrate a fairly common occurrence when making knockout animals: Null mice often die at some particular stage of development. This occurrence reveals that the disrupted gene is essential to cell function, and, at least during development of the organism, that failure to express its protein product is lethal to the organism. This might not surprise you, since you might readily hypothesize that loss of a gene that is so important to an essential process such as DNA replication would cause death of an organism. Scientists have the same expectations, and time and again knockout animals have confirmed this hypothesis.

The problem of lethality when creating knockouts led geneticists to develop **conditional knockout** approaches. (discussed in Chapter 20) This method allows one to eliminate a particular gene in a specific organ instead of the whole animal. This way the animal survives and researchers can study gene function in certain cell types. Conditional knockout animals can also be used to turn gene expression back on for rescue experiments.

References:

Adachi, N., et al. (2001). DNA Ligase IV-Deficient Cells Are More Resistant to Ionizing Radiation in the Absence of Ku70: Implications for DNA Double-strand Break Repair. *Proc. Natl. Acad. Sci. USA* 98:12109–12113.

Bentley, D.J., et al. (2002). DNA Ligase I Null Mouse Cells Show Normal DNA Repair Activity but Altered DNA Replication and Reduced Genome Stability. *J. Cell Sci.* 115:1551–1561.

Puebla-Osorio, N., et al. (2006). Early Embryonic Lethality due to Targeted Inactivation of DNA Ligase III. *Mol. Cell. Biol.* 26:3935–3941.

Questions to Consider:

1. Sometimes making null mice for a gene thought to have important functions results in animals with *no obvious mutant phenotype*. What conclusions might you reach about a gene that produces these results when a knockout is made?
2. How do you think it is possible for mice to develop for nine days without ligase 1—the ligase that is the “key ligase involved in DNA replication in mammalian cells”?
3. This *Modern Approaches to Understanding Gene Function* feature describes the role of three ligase genes, *Lig1, Lig3*, and *Lig4*, and their protein products. What happened to the mammalian *Lig2*? To find out, go to http://www.ncbi.nlm.nih.gov/ and type in “Human Lig2” using the “Gene” database. Look at your best hit. You may need to click on a replacement record to find the answer! Read the summary to find the solution to this mystery.

11.6 Eukaryotic DNA Replication Is Similar to Replication in Bacteria, but Is More Complex

Eukaryotic DNA replication shares many features with replication in bacteria. In both systems, double-stranded DNA is unwound at replication origins, replication forks are formed, and bidirectional DNA synthesis creates leading and lagging strands from single-stranded DNA templates under the direction of DNA polymerase. Eukaryotic polymerases have the same fundamental requirements for DNA synthesis as do bacterial polymerases: four deoxyribonucleoside triphosphates, a template, and a primer. However, eukaryotic DNA replication is more complex, due to several features of eukaryotic DNA. Eukaryotic cells contain much more DNA, this DNA is complexed with nucleosomes, and eukaryotic chromosomes are linear rather than circular. In this section, we will describe some of the ways in which eukaryotes deal with this added complexity.

Initiation at Multiple Replication Origins

The most obvious difference between eukaryotic and bacterial DNA replication is that eukaryotic replication must deal with greater amounts of DNA. For example, yeast cells contain 3 times as much DNA, and *Drosophila* cells contain 40 times as much as *E. coli* cells. In addition, eukaryotic DNA polymerases synthesize DNA at a rate 25 times slower (about 2000 nucleotides per minute) than that in bacteria. Under these conditions, replication from a single origin on a typical eukaryotic chromosome would take days to complete. However, replication of entire eukaryotic genomes is usually accomplished in a matter of minutes to hours.

To facilitate the rapid synthesis of large quantities of DNA, eukaryotic chromosomes contain multiple replication origins. Yeast genomes contain between 250 and 400 origins, and mammalian genomes have as many as 25,000. Multiple origins are visible under the electron microscope as "replication bubbles" that form as the DNA helix opens up, each bubble providing two potential replication forks (Figure 11.14). Origins in yeast, called **autonomously replicating sequences (ARSs),** consist of approximately 120 base pairs containing a **consensus sequence** (meaning a sequence that is the same, or nearly the same, in all yeast ARSs) of 11 base pairs. Origins in mammalian cells appear to be unrelated to specific sequence motifs and may be defined more by chromatin structure over a 6–55 kb region.

Eukaryotic replication origins not only act as sites of replication initiation, but also control the timing of DNA replication. These regulatory functions are carried out by a complex of more than 20 proteins, called the **prereplication complex (pre-RC),** which assembles at replication origins. In the early G1 phase of the cell cycle, replication origins are recognized by a six-protein complex known as an **origin recognition complex (ORC),** which tags the origin as a site of initiation of replication. Throughout the G1 phase of the cell cycle, other proteins associate with the ORC to form the pre-RC. The presence of a pre-RC at an origin "licenses" that origin for replication. Once DNA polymerases initiate synthesis at the origin, the pre-RC is disrupted and does not reassemble again until the G1 phase of the next cell cycle. This is an important mechanism because it distinguishes segments of DNA that have completed replication from segments of unreplicated DNA, thus maintaining orderly and efficient replication. It ensures that replication occurs only once along each stretch of DNA during each cell cycle.

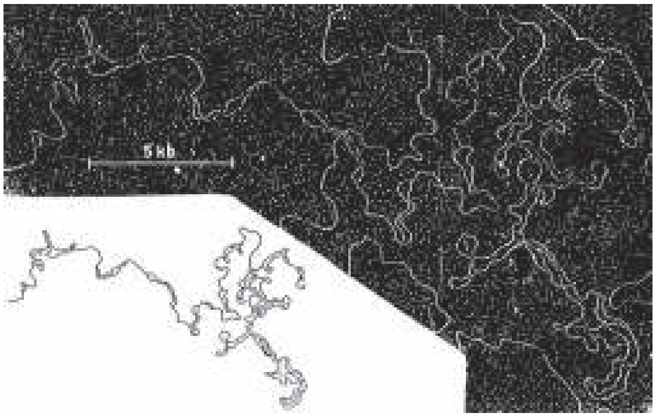


FIGURE 11.14 Demonstration of multiple origins of replication along a eukaryotic chromosome. DNA was isolated from a *Drosophila* nucleus and examined by electron microscopy. The DNA in this 119-kb fragment contains 23 replication bubbles, which can be located with the help of the diagram in the lower left-hand side of the electron micrograph.

The initiation of DNA replication is also regulated at the pre-RC. A number of cell-cycle kinases that phosphorylate replication proteins, along with helicases that unwind DNA, associate with the pre-RC and are essential for initiation. The kinases are activated in the S phase, at which time they phosphorylate other proteins that trigger the initiation of DNA replication. The end result is the unwinding of DNA at the replication forks, the stabilization of single-stranded DNA, the association of DNA polymerases with the origins, and the initiation of DNA synthesis.

Multiple Eukaryotic DNA Polymerases

To accommodate the large number of replicons, eukaryotic cells contain many more DNA polymerase molecules than do bacterial cells. For example, a single *E. coli* cell contains about 15 molecules of DNA polymerase III, but a mammalian cell contains tens of thousands of DNA polymerase molecules.

Eukaryotes also utilize a larger number of different DNA polymerase types than do bacteria. The human genome contains genes that encode at least 14 different DNA polymerases, only three of which are involved in the majority of nuclear genome DNA replication. The nomenclature and characteristics of these DNA polymerases are summarized in **Table 11.5**.

Pol α, δ, and ϵ are the major forms of the enzyme involved in initiation and elongation during eukaryotic nuclear DNA synthesis, so we will concentrate our discussion on these. Two of the four subunits of the **Pol α enzyme** synthesize RNA primers on both the leading and lagging strands. After the RNA primer reaches a length of about

TABLE 11.5 Properties of Eukaryotic DNA Polymerases

DNA Polymerase	Functions
α (alpha)	RNA/DNA primers, initiation of DNA synthesis
δ (delta)	Lagging strand synthesis, repair, recombination, proofreading
ϵ (epsilon)	Leading strand synthesis, repair, recombination, proofreading
γ (gamma)	Mitochondrial DNA replication and repair
β (beta)	Base-excision DNA repair
η (eta), ζ (zeta), κ (kappa), ν (nu), and ι (iota)	Translesion DNA synthesis
θ (theta), λ (lambda), μ (mu), ν (nu), and Rev1	DNA repair

10 ribonucleotides, another subunit adds 10–20 complementary deoxyribonucleotides. Pol α is said to possess low **processivity**, a term that refers to the strength of the association between the enzyme and its substrate, and thus the length of DNA that is synthesized before the enzyme dissociates from the template. Once the primer is in place, an event known as **polymerase switching** occurs, whereby Pol α dissociates from the template and is replaced by Pol δ or ϵ. These enzymes extend the primers on opposite strands of DNA, possess much greater processivity, and exhibit 3′ to 5′ exonuclease activity, thus having the potential to proofread. Pol ϵ synthesizes DNA on the leading strand, and Pol δ synthesizes the lagging strand. Both Pol δ and ϵ participate in other DNA synthesizing events in the cell, including several types of DNA repair and recombination. All three DNA polymerases are essential for viability.

As in bacterial DNA replication, the final stages in eukaryotic DNA replication involve replacing the RNA primers with DNA and ligating the Okazaki fragments on the lagging strand. In eukaryotes, the Okazaki fragments are about ten times smaller (100 to 150 nucleotides) than in bacteria.

Included in the remainder of DNA-replicating enzymes is Pol γ, which is found exclusively in mitochondria, synthesizing the DNA present in that organelle. Other DNA polymerases are involved in DNA repair and replication through regions of the DNA template that contain damage or distortions (called **translesion synthesis,** or **TLS**). Although these translesion DNA polymerases are less faithful in copying DNA and thus make more errors than Pol α, δ, and ϵ, they are able to bypass the distortions, leaving behind gaps that may then be repaired. We will return to the topic of DNA damage and repair later in the text (see Chapter 15).

Replication through Chromatin

One of the major differences between bacterial and eukaryotic DNA is that eukaryotic DNA is complexed with DNA-binding proteins, existing in the cell as *chromatin*. As we will discuss later in the text (see Chapter 12), chromatin consists of regularly repeating units called nucleosomes, each of which consists of about 200 base pairs of DNA wrapped around eight histone protein molecules. Before DNA polymerases can begin synthesis, nucleosomes and other DNA-binding proteins must be stripped away or otherwise modified to allow the passage of replication proteins. As DNA synthesis proceeds, the histones and nonhistone proteins must rapidly reassociate with the newly formed duplexes, reestablishing the characteristic nucleosome pattern. Electron microscopy studies, such as the one shown in **Figure 11.15**, show that nucleosomes form immediately after new DNA is synthesized at replication forks.

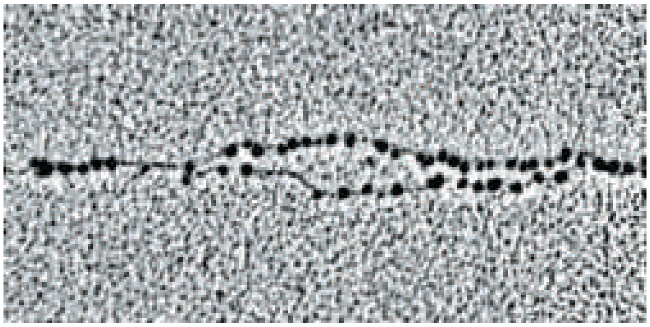


FIGURE 11.15 An electron micrograph of a eukaryotic replicating fork demonstrating the presence of histone-protein-containing nucleosomes on both branches.

In order to re-create nucleosomal chromatin on replicated DNA, the synthesis of new histone proteins is tightly coupled to DNA synthesis during the S phase of the cell cycle. Research data suggest that nucleosomes are disrupted just ahead of the replication fork and that the preexisting histone proteins can assemble with newly synthesized histone proteins into new nucleosomes. The new nucleosomes are assembled behind the replication fork, onto the two daughter strands of DNA. The assembly of new nucleosomes is carried out by **chromatin assembly factors (CAFs)** that move along with the replication fork.

11.7 Telomeres Solve Stability and Replication Problems at Eukaryotic Chromosome Ends

A final difference between bacterial and eukaryotic DNA synthesis stems from the structural differences in their chromosomes. Unlike the closed, circular DNA of bacteria and most bacteriophages, eukaryotic chromosomes are linear. The presence of linear DNA “ends” on eukaryotic chromosomes creates two potential problems.

The first problem is that the double-stranded “ends” of DNA molecules at the termini of linear chromosomes resemble the **double-stranded breaks (DSBs)** that can occur when a chromosome becomes broken internally as a result of DNA damage. Such double-stranded DNA ends are recognized by the cell’s DNA repair mechanisms that join the “loose ends” together, leading to chromosome fusions and translocations. If the ends do not fuse, they are vulnerable to degradation by nucleases. The second problem occurs during DNA replication, because DNA polymerases cannot synthesize new DNA at the tips of single-stranded 5′ ends.

To deal with these two problems, linear eukaryotic chromosomes end in distinctive sequences called telomeres, as we will describe next.

Telomere Structure and Chromosome Stability

In 1978, Elizabeth Blackburn and Joe Gall reported the presence of unexpected structures at the ends of chromosomes of the ciliated protozoan *Tetrahymena*. They showed that the protozoan's chromosome ends consisted of the short sequence 5′-TTGGGG-3′, tandemly repeated from 30 to 60 times. This strand is referred to as the G-rich strand, in contrast to its complementary strand, the so-called C-rich strand, which displays the repeated sequence 5′-AACCCC-3′. Since then, researchers have discovered similar tandemly repeated DNA sequences at the ends of linear chromosomes in most eukaryotes. These repeat regions make up the chromosome's **telomeres.** In humans, the telomeric sequence 5′-TTAGGG-3′ is repeated several thousand times and telomeres can vary in length from 5 to 15 kb. In contrast, yeast telomeres are several 100 base pairs long and mouse telomeres are between 20 and 50 kb long. Since each linear chromosome ends with two DNA strands running antiparallel to one another, one strand has a 3′ ending and the other has a 5′ ending. It is the 3′ strand that is the G-rich one. This has special significance during telomere replication.

Two features of telomeric DNA help to explain how telomeres protect the ends of linear chromosomes. First, at the end of telomeres, a stretch of single-stranded DNA extends out from the 3′ G-rich strand. This single-stranded tail varies in length between organisms. In *Tetrahymena*, the tail is between 12 and 16 nucleotides long, whereas in mammals, it varies between 30 and 400 nucleotides long. The 3′ ends of G-rich single-stranded tails are capable of interacting with upstream sequences within the tail, creating loop structures. The loops, called **t-loops,** resemble those created when you tie your shoelaces into a bow. Second, a complex of six proteins binds and stabilizes telomere t-loops, forming the **shelterin complex.** It is believed that t-loop structures, in combination with the shelterin proteins, close off the ends of chromosomes and make them resistant to nuclease digestion and DNA fusions. Shelterin proteins also help to recruit telomerase enzymes to telomeres during telomere replication, which we discuss next.

Telomeres and Chromosome End Replication

Now let's consider the problem that semiconservative replication poses the ends of double-stranded DNA molecules. As we have learned previously in this chapter, DNA replication initiates from short RNA primers, synthesized on both leading and lagging strands (Figure 11.16). Primers are necessary because DNA polymerase requires a free 3′-OH on which to initiate synthesis. After replication is completed, these RNA primers are removed. The resulting gaps within the new daughter strands are filled by DNA polymerase and sealed by ligase. These internal gaps have free 3′-OH groups available at the ends of the Okazaki fragments for DNA polymerase to initiate synthesis. The problem arises at the gaps left at the 5′ ends of the newly synthesized DNA [gaps (b) and (c) in Figure 11.16]. These gaps cannot be filled by DNA polymerase because no free 3′-OH groups are available for the initiation of synthesis.

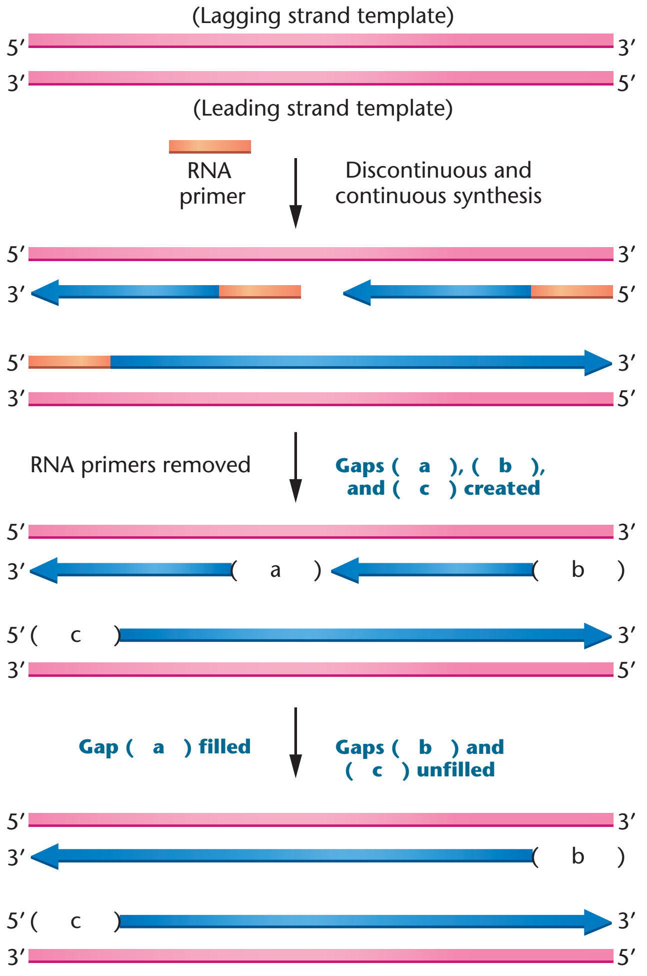

(Lagging strand template)
5′
3′
3′
5′
(Leading strand template)
RNA primer
Discontinuous and continuous synthesis
5′
3′
3′
5′
5′
3′
3′
5′
RNA primers removed
Gaps (a), (b), and (c) created
5′
3′
(a)
(b)
5′ (c)
3′
3′
5′
Gap (a) filled
Gaps (b) and (c) unfilled
5′
3′
(b)
5′ (c)
3′
3′
5′

FIGURE 11.16 Diagram illustrating the difficulty encountered during the replication of the ends of linear chromosomes. Of the three gaps, (b) and (c) are left following synthesis on both the leading and lagging strands.

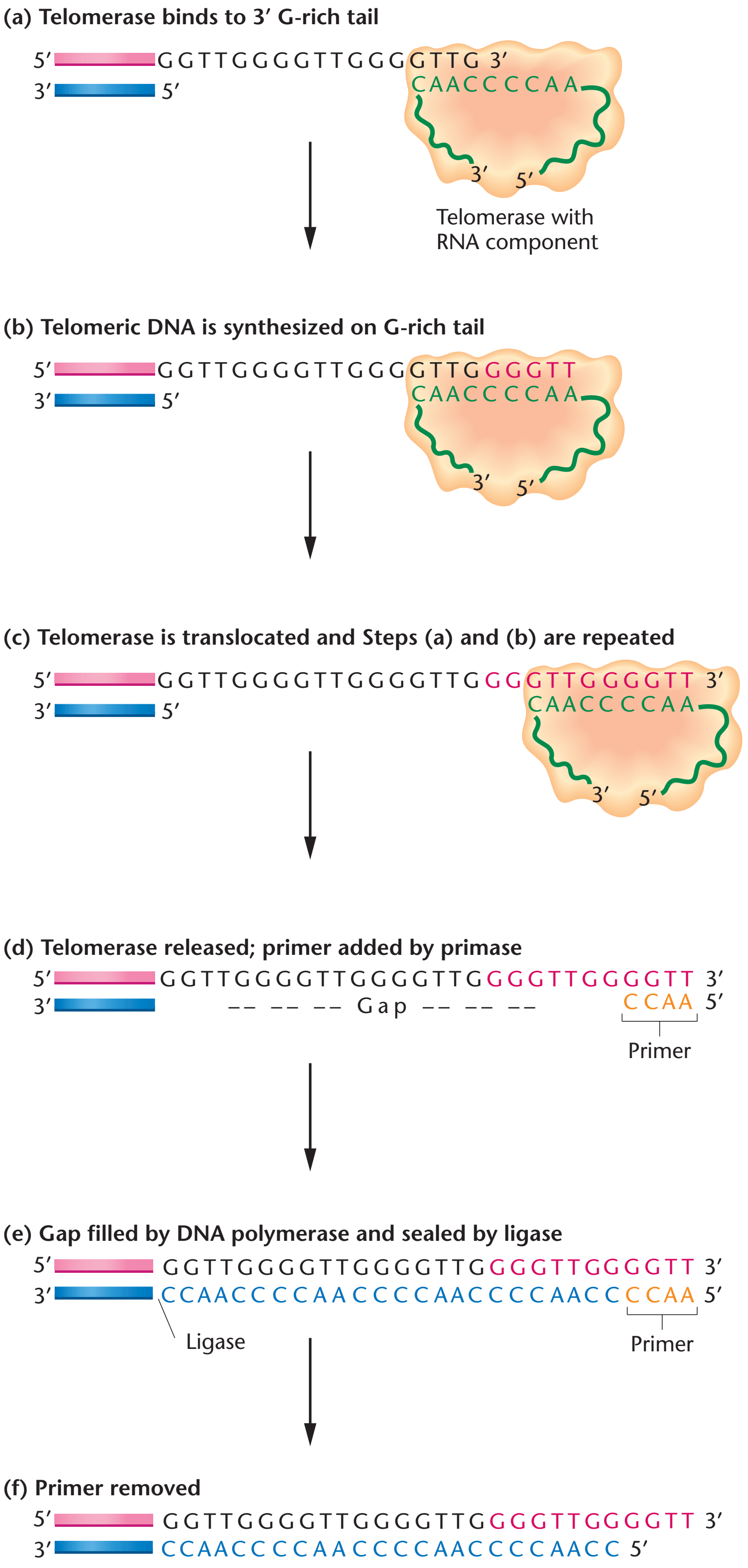
Thus, in the situation depicted in Figure 11.16, gaps remain on newly synthesized DNA strands at each successive round of synthesis, shortening the double-stranded ends of the chromosome by the length of the RNA primer. With each round of replication, the shortening becomes more severe in each daughter cell, eventually extending beyond the telomere and potentially deleting gene-coding regions.

The solution to this so-called **end-replication problem** is provided by a unique eukaryotic enzyme called **telomerase.** Telomerase was first discovered by Elizabeth Blackburn and her graduate student, Carol Greider, in studies of *Tetrahymena*. As noted earlier, telomeric DNA in eukaryotes consists of many short, repeated nucleotide sequences, with the G-rich strand overhanging in the form of a single-stranded tail. In *Tetrahymena* the tail contains several repeats of the sequence 5′-TTGGGG-3′. As we will see, telomerase is capable of adding several more repeats of this six-nucleotide sequence to the 3′end of the G-rich strand. Detailed investigation by Blackburn and Greider of how the *Tetrahymena* telomerase enzyme accomplishes this synthesis yielded an extraordinary finding. The enzyme is a *ribonucleoprotein,* containing within its molecular structure a short piece of RNA that is essential to its catalytic activity. The **telomerase RNA component (TERC)** serves as both a "guide" to proper attachment of the enzyme to the telomere and a "template" for synthesis of its DNA complement. Synthesis of DNA using RNA as a template is called **reverse transcription**. The **telomerase reverse transcriptase (TERT)** is the catalytic subunit of the telomerase enzyme. In addition to TERC and TERT, telomerase contains a number of accessory proteins. In *Tetrahymena,* TERC contains the sequence CAACCCCAA, within which is found the complement of the repeating telomeric DNA sequence that must be synthesized (TTGGGG).

Figure 11.17 shows a model of how researchers envision the enzyme working. Part of the TERC RNA sequence of the enzyme (shown in green) base-pairs with the ending sequence of the single-stranded overhanging DNA, while the remainder of the TERC RNA extends beyond the overhang. Next, the telomerase's reverse transcription activity synthesizes a stretch of single-stranded DNA using the TERC RNA as a template, thus increasing the length of the 3′ tail. It is believed that the enzyme is then translocated toward the (newly formed) end of the tail, and the same events are repeated, continuing the extension process.

Once the telomere 3′ tail has been lengthened by telomerase, conventional DNA synthesis ensues. Primase lays down a primer near the end of the telomere tail; then DNA polymerase and ligase fill most of the gap [**Figure 11.17(d)** and **(e)**]. When the primer is removed, a small gap remains at the end of the telomere. However, this gap is located well beyond the original end of the chromosome, thus preventing any chromosome shortening.

Telomerase function has now been found in all eukaryotes studied. As we will discuss later in the text (see Chapter 12), telomeric DNA sequences have been highly conserved


(a) Telomerase binds to 3′ G-rich tail

(b) Telomeric DNA is synthesized on G-rich tail

(c) Telomerase is translocated and Steps (a) and (b) are repeated

(d) Telomerase released; primer added by primase

(e) Gap filled by DNA polymerase and sealed by ligase

(f) Primer removed

5′ GGTTGGGGTTGGGGTTG 3′
3′ 5′
CAACCCCAA
3′ 5′
Telomerase with RNA component
5′ GGTTGGGGTTGGGGTTGGGGTT
3′ 5′
CAACCCCAA
3′ 5′
5′ GGTTGGGGTTGGGGTTGGGGTTGGGGTT 3′
3′ 5′
CAACCCCAA
3′ 5′
5′ GGTTGGGGTTGGGGTTGGGGTTGGGGTT 3′
3′ -- -- -- Gap -- -- -- CCAA 5′
Primer
5′ GGTTGGGGTTGGGGTTGGGGTTGGGGTT 3′
3′ CCAACCCCAACCCCAACCCCAACCCCAA 5′
Ligase
Primer
5′ GGTTGGGGTTGGGGTTGGGGTTGGGGTT 3′
3′ CCAACCCCAACCCCAACCCCAACC 5′

FIGURE 11.17 The predicted solution to the problem posed in Figure 11.16. The enzyme telomerase (with its TERC RNA component shown in green) directs synthesis of repeated TTGGGG sequences, resulting in the formation of an extended 3′ overhang. This facilitates DNA synthesis on the opposite strand, filling in the gap that would otherwise have been increased in length with each replication cycle.

throughout evolution, reflecting the critical function of telomeres.

Telomeres in Disease, Aging, and Cancer

Despite the importance of maintaining telomere length for chromosome integrity, only some cell types express telomerase. In humans, these include embryonic stem cells, some types of adult stem cells, and other cell types that need to divide repeatedly such as epidermal cells and cells of the immune system. In contrast, most normal somatic cells do not express telomerase. As a result, after many cell divisions, somatic cell telomeres become seriously eroded, leading to chromosome damage that can either kill the cell or cause it to cease dividing and enter a state called senescence.

Several rare human diseases have been associated with loss of telomerase activity and abnormally short telomeres. For example, patients with the inherited form of dyskeratosis congenita have mutations in genes encoding telomerase or shelterin subunits. These mutations bring about many different symptoms that are also seen in premature aging, and patients suffer early deaths due to stem cell failure. Some premature aging syndromes, such as Fanconi anemia and Werner syndrome are associated with short telomeres; however, the mutated genes in these syndromes are those involved in DNA damage repair and not telomere maintenance. Whether telomere shortening itself is causative of, or the result of, these mutations is not clear. Many studies show a correlation between telomere length or telomerase activity and common diseases such as diabetes or heart disease.

The connection between telomere length and aging has been the subject of much research and speculation. As we discussed previously, when telomeres become critically short, a cell may suffer chromosome damage and enter **senescence,** a state in which cell division ceases and the cell undergoes metabolic changes that cause it to function less efficiently. Some scientists propose that the presence of senescent cells in a multicellular organism may bring about the physiological changes associated with aging. The topic of telomeres and aging is discussed in the "Genetics, Ethics, and Society" essay later in this chapter.

Although normal somatic cells contain little if any telomerase, shorten their telomeres, and undergo senescence after multiple cell divisions, cancer cells do not. More than 90 percent of human cancer cells contain telomerase activity and maintain telomere length though many cell divisions—becoming "immortal." Those cancer cells that do not contain telomerase use a different telomere-lengthening method called *alternative lengthening of telomeres (ALT)*. ALT uses homologous recombination mechanisms to maintain their telomeres. (Homologous recombination is discussed in Section 11.8.) In tissue cultures, cells can be transformed into cancer cells by introducing telomerase activity, as long as at least two other types of genes (proto-oncogenes and tumor-suppressor genes) are mutated or abnormally expressed. By itself, telomerase activity is not sufficient to create a cancer cell. We return to the multistep nature of cancer genetics later (see Chapter 24).

The requirement for telomerase activity in cancer cells suggests that researchers may be able to develop cancer drugs that repress tumor growth by inhibiting telomerase activity. Because most human somatic cells do not express telomerase, such a therapy might be relatively tumor specific and less toxic than many current anticancer drugs. A number of such anti-telomerase drugs are currently being developed, and some have entered phase III clinical trials.

11.8 Recombination Is Essential for Genetic Exchange and DNA Repair

We now turn to a topic previously discussed (see Chapters 2 and 5). We were introduced to genetic crossing over, a phenomenon that occurs in eukaryotes during prophase I of meiosis. During crossing over, members of homologous chromosomes pair with each other and exchange chromosome segments, resulting in chromosomes that are combinations of genetic information from paternal and maternal sources. This genetic exchange is a major source of genetic diversity through sexual reproduction. We saw (in Chapter 6) the various ways in which genetic information can be exchanged between DNA molecules from bacteria or bacteriophages. In these examples of recombination, the process of exchange begins with either single-stranded or double-stranded DNA breaks.

Not only is recombination essential for the exchange of genetic information during meiosis, but it is also important for the repair of damaged DNA. As we discuss later (see Chapter 15), various types of agents can create double-stranded breaks in DNA molecules. These agents include ionizing radiation and free radicals. DNA recombination is also part of the ALT mechanism used to maintain eukaryotic telomeres, as discussed in Section 11.7. When cells are defective in recombination mechanisms, they are prone to accumulating mutations in genes and chromosomes, which in turn may lead to the development of tumors, diseases, or even cell death.

Now that we have discussed the chemistry and replication of DNA, we can consider how recombination occurs at the molecular level. In general, our discussion pertains to genetic exchange between any two homologous double-stranded DNA molecules, whether they are viral or bacterial chromosomes or eukaryotic homologs during meiosis.

Models of Homologous Recombination

Genetic exchange at equivalent positions along two chromosomes with substantial DNA sequence homology is referred to as **general,** or **homologous, recombination.**

Several models attempt to explain homologous recombination, but they all have certain features in common. First, all are based on proposals first put forth independently by Robin Holliday and Harold L. K. Whitehouse in 1964. Second, they all depend on the complementarity between DNA strands to explain the precision of the exchange. Finally, each model relies on a series of enzymatic processes in order to accomplish genetic recombination.

One such model, beginning from a single-stranded DNA break, is shown in **Figure 11.18**. It begins with two paired

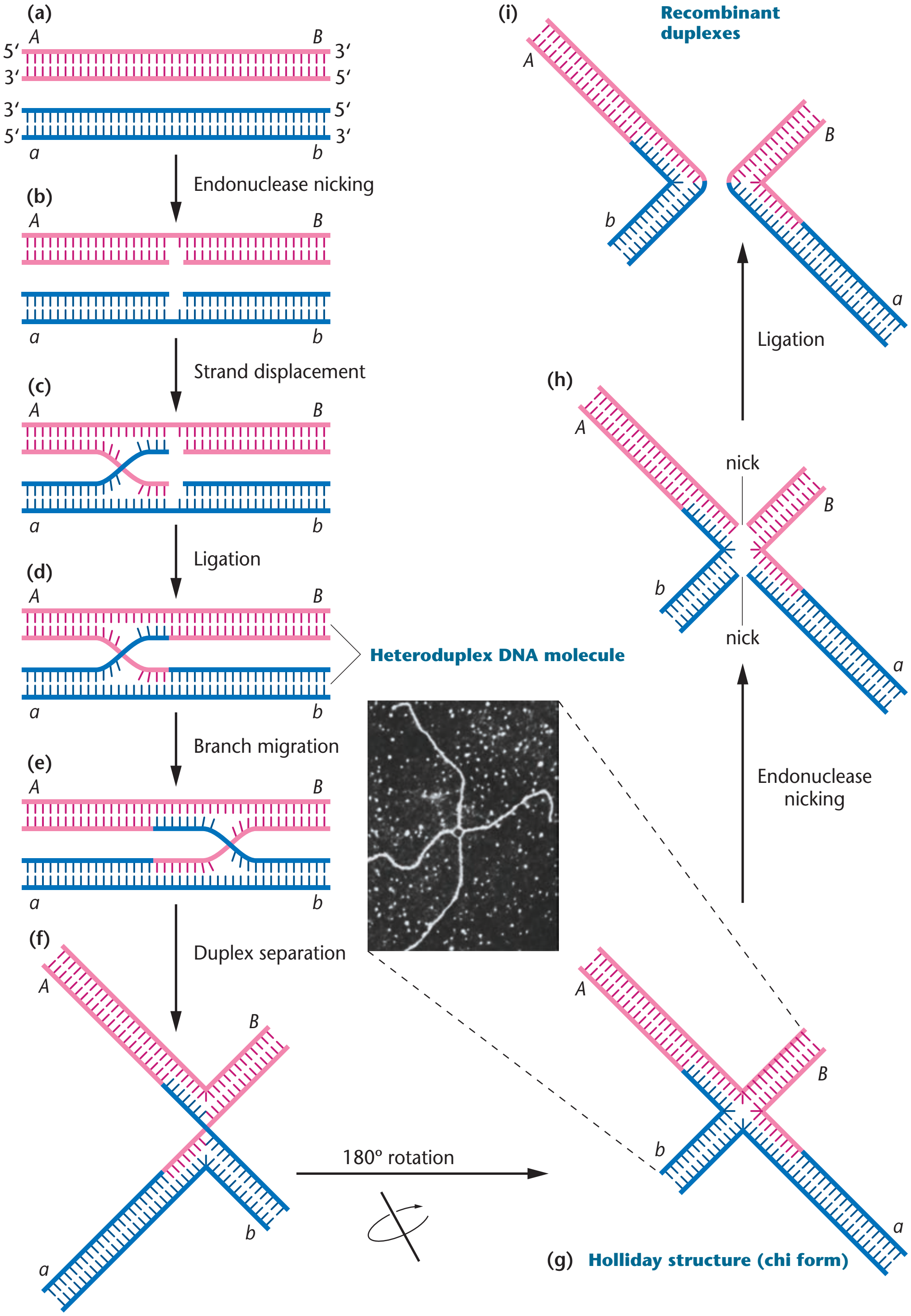


FIGURE 11.18 Model depicting how genetic recombination can occur as a result of the breakage and rejoining of heterologous DNA strands. Each stage is described in the text. The electron micrograph shows DNA in a χ-form structure similar to the diagram in (g); the DNA is an extended Holliday structure, derived from the *Col*E1 plasmid of *E. coli.*

DNA duplexes, or homologs [Step (a)], in each of which an endonuclease introduces a single-stranded nick at an identical position [Step (b)]. The internal strand endings produced by these cuts are then displaced and subsequently pair with their complements on the opposite duplex [Step (c)]. Next, a ligase seals the loose ends [Step (d)], creating hybrid duplexes called **heteroduplex DNA molecules,** held together by a cross-bridge structure. The position of this cross bridge can then move down the chromosome by a process referred to as **branch migration** [Step (e)], which occurs as a result of a zipper-like action as hydrogen bonds are broken and then reformed between complementary bases of the displaced strands of each duplex. This migration yields an increased length of heteroduplex DNA on both homologs.

If the duplexes bend [Step {f}] and the bottom portion shown in the figure rotates 180° [Step (g)], an intermediate planar structure called a χ (chi) form—or **Holliday structure**—is created. If the two strands on opposite homologs previously uninvolved in the exchange are now nicked by an endonuclease [Step (h)] and ligation occurs as in Step (i), two recombinant duplexes are created. Note that the arrangement of alleles is altered as a result of this recombination.

Whereas the preceding model involves single-stranded breaks, similar recombination models have been proposed that involve double-stranded breaks in one of the DNA double helices. In these models, endonucleases remove nucleotides at the breakpoint, creating 3′ overhangs on each strand. One of the broken strands invades the intact double helix of the other homolog, and both strands line up with the intact homolog. DNA repair synthesis then fills all gaps, and two Holliday junctions are formed. Endonuclease cleavages and ligations finalize the exchange. The end result is the same as our original model: genetic exchange occurs between homologous DNA molecules.

GENETICS, ETHICS, AND SOCIETY

Telomeres: The Key to a Long Life?

We humans, like all multicellular organisms, grow old and die. As we age, our immune systems become less efficient, wound healing is impaired, and tissues lose resilience. Why do we go through these age-related declines, and can we reverse the march to mortality? Some researchers suggest that the answers to these questions may lie at the ends of our chromosomes.

Human cells, both those in our bodies and those growing in culture dishes, have a finite life span. When placed into tissue culture dishes, normal human fibroblasts become senescent, losing their ability to grow and divide after about 50 cell divisions. Eventually, they die. Although we don't know whether cellular senescence directly causes aging in a multicellular organism, the evidence is suggestive. For example, cultured cells derived from young people undergo more divisions than those from older people; cells from short-lived species stop growing after fewer divisions than those from longer-lived species; and cells from patients with premature aging syndromes undergo fewer divisions than those from normal patients.

One of the many characteristics of aging cells involves telomeres. As described in this chapter, most mammalian somatic cells do not contain telomerase activity and telomeres shorten with each DNA replication. Some epidemiological studies show a correlation between telomere length in humans and their life spans. In addition, some common diseases such as cardiovascular disease and some lifestyle factors such as smoking, poor diets, and stress, correlate with shorter than average telomere lengths. Despite these correlations, the data linking telomere length and longevity in humans are not consistent and remain the subject of scientific debate.

Telomerase activity has also been correlated with aspects of aging in multicellular organisms. In one study, investigators introduced cloned telomerase genes into normal human cells in culture. The increase in telomerase activity caused the telomeres' lengths to increase, and the cells continued to grow past their typical senescence point. In another study, researchers created a strain of mice that was defective in the TERT subunit of telomerase. These mice developed extremely short telomeres and showed the classic symptoms of aging, including tissue atrophy, neurodegeneration, and a shortened life span. When the researchers reactivated telomerase function in these prematurely aging adult mice, tissue atrophies and neurodegeneration were reversed and their life spans increased. Similar studies led to the conclusion that overexpression of telomerase in normal mice also increased their life spans, although it was not clear that telomere lengths were altered. These studies suggest that some of the symptoms that accompany old age in humans might be reversed by activating telomerase genes. However, some scientists still debate whether telomerase activation or telomere lengthening directly cause these effects or may simply accompany other, unknown, causative mechanisms.

YOUR TURN

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand some of the research and ethical questions that surround the links between telomeres and aging.

1. The connection between telomeres and aging has been of great interest to both scientists and the public. Studies have used model organisms, cultured cells, and data from epidemiological surveys to try to determine a correlation. Although these studies suggest links between telomeres and aging, the conclusions from these studies have also been the subject of debate. How would you assess the current status of research on the link between telomeres and human aging?

   Begin your exploration of current telomere research with two reviews that draw differing conclusions from the same data: Bär, C. and Blasco, M. A. (2016). Telomeres and Telomerase as Therapeutic Targets to Prevent and Treat Age-Related Diseases. *F1000Res* 5 (F1000 Faculty Rev): 89, and Simons, M. J. P. (2015). Questioning Causal Involvement of Telomeres in Aging. *Aging Res. Rev.* 24:191–196.

2. A large number of private companies now offer telomere length diagnostic tests and telomere lengthening supplements to the public. Discuss the ethics of offering such tests and supplements when some scientists argue that it is too early to sufficiently understand the mechanisms that may (or may not) link telomere length or telomerase activity with aging. What are the potential benefits and harms of these tests and treatments? Would you purchase these tests or supplements? Why or why not?

   This topic is discussed in Leslie, M. (2011). Are Telomere Tests Ready for Prime Time? *Science* 332:414–415.

CASE STUDY At loose ends

Dyskeratosis congenita (DKC) is a rare human genetic disorder affecting telomere replication. Mutations in the genes encoding the protein or RNA subunits of telomerase result in very short telomeres. DKC symptoms include bone marrow failure (reduced production of blood cells) and anemia. If symptoms are severe, a bone marrow transplant may be the only form of effective treatment. In one case, clinicians recommended that a 27-year-old woman with a dominant form of DKC undergo a bone marrow transplant to treat the disorder. Her four siblings were tested, and her 13-year-old brother was identified as the best immunologically matched donor. However, before being tested, he was emphatic that he did not want to know if he had DKC. During testing, it was discovered that he had unusually short telomeres and would most likely develop symptoms of DKC.

1. Why might mutations in genes encoding telomerase subunits lead to bone marrow failure?
2. Although the brother is an immunologically matched donor for his sister, it would be unethical for the clinicians to transplant bone marrow from the brother to the sister. Why?
3. The clinicians are faced with another ethical dilemma. How can they respect the brother's desire to not know if he has DKC while also revealing that he is not a suitable donor for his sister? In addition, what should the clinicians tell the sister about her brother?

See Denny, C., et al. (2008). All in the Family: Disclosure of "Unwanted" Information to an Adolescent to Benefit a Relative. *Am. J. Med. Genet.* 146A(21):2719–2724.

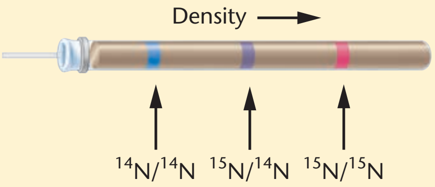
Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

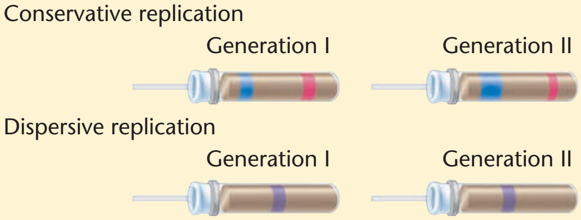
1. In 1958, Meselson and Stahl resolved the question of which of three potential modes of replication is utilized by *E. coli* during the duplication of DNA in favor of semiconservative replication, showing that newly synthesized DNA consists of one old strand and one new strand.
2. Taylor, Woods, and Hughes demonstrated semiconservative replication in eukaryotes using the root tips of the broad bean as the source of dividing cells.
3. Arthur Kornberg isolated the enzyme DNA polymerase I from *E. coli* and showed that it is capable of directing *in vitro* DNA synthesis, provided that a template and precursor nucleoside triphosphates are supplied.
4. The discovery of the *polA1* mutant strain of *E. coli,* capable of DNA replication despite its lack of polymerase I activity, cast doubt on the enzyme's hypothesized *in vivo* replicative function. Polymerase III has been identified as the enzyme responsible for DNA replication *in vivo.*
5. During the initiation of DNA synthesis, the double helix unwinds, forming a replication fork at which synthesis begins. Proteins stabilize the unwound helix and assist in relaxing the coiling tension created ahead of the replication fork.
6. DNA synthesis is initiated at specific sites along each template strand by the enzyme primase, resulting in short segments of RNA that provide suitable 3′ ends upon which DNA polymerase III can begin polymerization.
7. Concurrent DNA synthesis occurs continuously on the leading strand and discontinuously on the opposite lagging strand, resulting in short Okazaki fragments that are later joined by DNA ligase.
8. DNA replication in eukaryotes is more complex than replication in bacteria, using multiple replication origins, multiple forms of DNA polymerases, and factors that disrupt and assemble nucleosomal chromatin.
9. Replication at the ends of linear chromosomes in eukaryotes poses a special problem that can be solved by the presence of telomeres and by a unique RNA-containing enzyme called telomerase.
10. Homologous recombination between DNA molecules is essential for the exchange of genetic information and the repair of damaged DNA.

INSIGHTS AND SOLUTIONS

1. Predict the theoretical results of conservative and dispersive replication of DNA under the conditions of the Meselson—Stahl experiment. Follow the results through two generations of replication after cells have been shifted to a ^{14}N-containing medium, using the following sedimentation pattern.



Solution:



2. Mutations in the *dnaA* gene of *E. coli* are lethal and can only be studied following the isolation of conditional, temperature-sensitive mutations. Such mutant strains grow nicely and replicate their DNA at the permissive temperature of 18°C, but they do not grow or replicate their DNA at the restrictive temperature of 37°C. Two observations were useful in determining the function of the DnaA protein product. First, *in vitro* studies using DNA templates that have unwound do not require the DnaA protein. Second, if intact cells are grown at 18°C and are then shifted to 37°C, DNA synthesis continues at this temperature until one round of replication is completed and then stops. What do these observations suggest about the role of the *dnaA* gene product?

Solution:
At 18°C (the permissive temperature), the mutation is not expressed and DNA synthesis begins. Following the shift to the restrictive temperature, the already initiated DNA synthesis continues, but no new synthesis can begin. Because the DnaA protein is not required for synthesis of unwound DNA, these observations suggest that, *in vivo,* the DnaA protein plays an essential role in DNA synthesis by interacting with the intact helix and somehow facilitating the localized denaturation necessary for synthesis to proceed.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. HOW DO WE KNOW? In this chapter, we focused on how DNA is replicated and synthesized. We also discussed recombination at the DNA level. Along the way, we encountered many opportunities to consider how this information was acquired. On the basis of these discussions, what answers would you propose to the following fundamental questions?
(a) What is the experimental basis for concluding that DNA replicates semiconservatively in both bacteria and eukaryotes?
(b) How was it demonstrated that DNA synthesis occurs under the direction of DNA polymerase III and not polymerase I?
(c) How do we know that *in vivo* DNA synthesis occurs in the 5′ to 3′ direction?
(d) How do we know that DNA synthesis is discontinuous on one of the two template strands?
(e) What observations reveal that a “telomere problem” exists during eukaryotic DNA replication, and how did we learn of the solution to this problem?

2. CONCEPT QUESTION Review the Chapter Concepts list on p. 238. These are concerned with the replication and synthesis of DNA. Write a short essay that distinguishes between the terms *replication* and *synthesis*, as applied to DNA. Which of the two is most closely allied with the field of biochemistry?

3. Compare conservative, semiconservative, and dispersive modes of DNA replication.

4. Describe the role of ^{15}N in the Meselson—Stahl experiment.

5. Predict the results of the experiment by Taylor, Woods, and Hughes if replication were (a) conservative and (b) dispersive.

6. What are the requirements for *in vitro* synthesis of DNA under the direction of DNA polymerase I?

7. In Kornberg’s initial experiments, it was rumored that he grew *E. coli* in Anheuser-Busch beer vats. (Kornberg was working at Washington University in St. Louis.) Why do you think this might have been helpful to the experiment?

8. How did Kornberg assess the fidelity of DNA polymerase I in copying a DNA template?

9. Which characteristics of DNA polymerase I raised doubts that its *in vivo* function is the synthesis of DNA leading to complete replication?

10. Kornberg showed that nucleotides are added to the 3′ end of each growing DNA strand. In what way does an exposed 3′-OH group participate in strand elongation?

11. What was the significance of the *polA1* mutation?

12. Summarize and compare the properties of DNA polymerase I, II, and III.

13. List and describe the function of the ten subunits constituting DNA polymerase III. Distinguish between the holoenzyme and the core enzyme.

14. Distinguish between (a) unidirectional and bidirectional synthesis, and (b) continuous and discontinuous synthesis of DNA.

15. List the proteins that unwind DNA during *in vivo* DNA synthesis. How do they function?

16. Define and indicate the significance of (a) Okazaki fragments, (b) DNA ligase, and (c) primer RNA during DNA replication.

17. Outline the current model for DNA synthesis.

18. Why is DNA synthesis expected to be more complex in eukaryotes than in bacteria? How is DNA synthesis similar in the two types of organisms?

19. Suppose that *E. coli* synthesizes DNA at a rate of 100,000 nucleotides per minute and takes 40 minutes to replicate its chromosome. (a) How many base pairs are present in the entire *E. coli* chromosome? (b) What is the physical length of the chromosome in its helical configuration—that is, what is the circumference of the chromosome if it were opened into a circle?

20. Several temperature-sensitive mutant strains of *E. coli* display the following characteristics. Predict what enzyme or function is being affected by each mutation.

(a) Newly synthesized DNA contains many mismatched base pairs.

(b) Okazaki fragments accumulate, and DNA synthesis is never completed.

(c) No initiation occurs.

(d) Synthesis is very slow.

(e) Supercoiled strands remain after replication, which is never completed.

21. While many commonly used antibiotics interfere with protein synthesis or cell wall formation, clorobiocin, one of several antibiotics in the aminocoumarin class, inhibits the activity of bacterial DNA gyrase. Similar drugs have been tested as treatments for human cancer. How might such drugs be effective against bacteria as well as cancer?

22. Describe the “end-replication problem” in eukaryotes. How is it resolved?

23. Many of the gene products involved in DNA synthesis were initially defined by studying mutant *E. coli* strains that could not synthesize DNA.

(a) The *dnaE* gene encodes the α subunit of DNA polymerase III. What effect is expected from a mutation in this gene? How could the mutant strain be maintained?

(b) The *dnaQ* gene encodes the ϵ subunit of DNA polymerase. What effect is expected from a mutation in this gene?

24. In 1994, telomerase activity was discovered in human cancer cell lines. Although telomerase is not active in human somatic tissue, human somatic cells do contain the genes for telomerase proteins and telomerase RNA. Since inappropriate activation of telomerase may contribute to cancer, why do you think the genes coding for this enzyme have been maintained in the human genome throughout evolution? Are there any types of human body cells where telomerase activation would be advantageous or even necessary? Explain.

Extra-Spicy Problems

25. The genome of *D. melanogaster* consists of approximately 1.7×10^8 base pairs. DNA synthesis occurs at a rate of 30 base pairs per second. In the early embryo, the entire genome is replicated in five minutes. How many bidirectional origins of synthesis are required to accomplish this feat?

26. Assume a hypothetical organism in which DNA replication is conservative. Design an experiment similar to that of Taylor, Woods, and Hughes that will unequivocally establish this fact. Using the format established in Figure 11.5, draw sister chromatids and illustrate the expected results establishing this mode of replication.

27. DNA polymerases in all organisms add only 5′ nucleotides to the 3′ end of a growing DNA strand, never to the 5′ end. One possible reason for this is the fact that most DNA polymerases have a proofreading function that would not be *energetically* possible if DNA synthesis occurred in the 3′ to 5′ direction.

(a) Sketch the reaction that DNA polymerase would have to catalyze if DNA synthesis occurred in the 3′ to 5′ direction.

(b) Consider the information in your sketch and speculate as to why proofreading would be problematic.

28. Assume that the sequence of bases shown below is present on one nucleotide chain of a DNA duplex and that the chain has opened up at a replication fork. Synthesis of an RNA primer occurs on this template starting at the base that is underlined.

(a) If the RNA primer consists of eight nucleotides, what is its base sequence?

(b) In the intact RNA primer, which nucleotide has a free 3′-OH terminus?

3′.......GGCTACC$\underline{\text{T}}$GGATTCA....5′

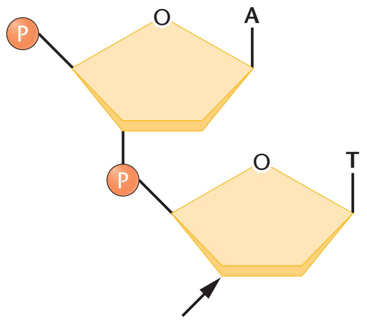
29. Reiji and Tuneko Okazaki conducted a now classic experiment in 1968 in which they discovered a population of short fragments synthesized during DNA replication. They introduced a short pulse of ^{3}H-thymidine into a culture of *E. coli* and extracted DNA from the cells at various intervals. In analyzing the DNA after centrifugation in denaturing gradients, they noticed that as the interval between the time of ^{3}H-thymidine introduction and the time of centrifugation increased, the proportion of short strands decreased and more labeled DNA was found in larger strands. What would account for this observation?

30. Consider the drawing of a dinucleotide below.

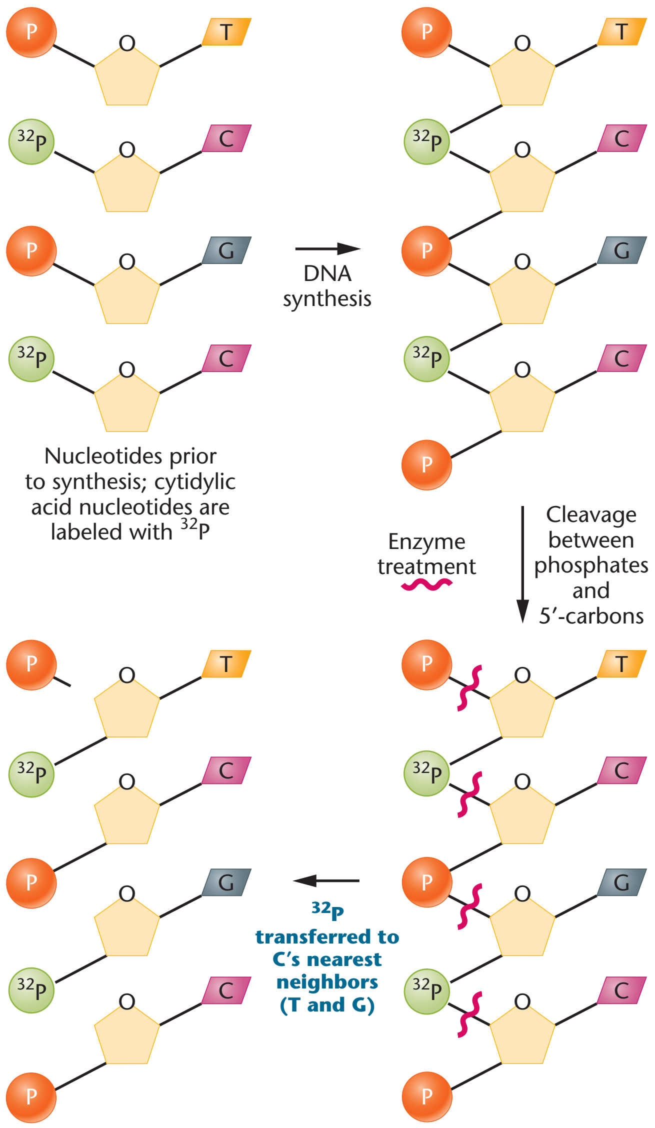
(a) Is it DNA or RNA?

(b) Is the arrow closest to the 5′ or the 3′ end?

(c) Suppose that the molecule was cleaved with the enzyme spleen phosphodiesterase, which breaks the covalent bond connecting the phosphate to C-5′. After cleavage, to which nucleoside is the phosphate now attached (A or T)?

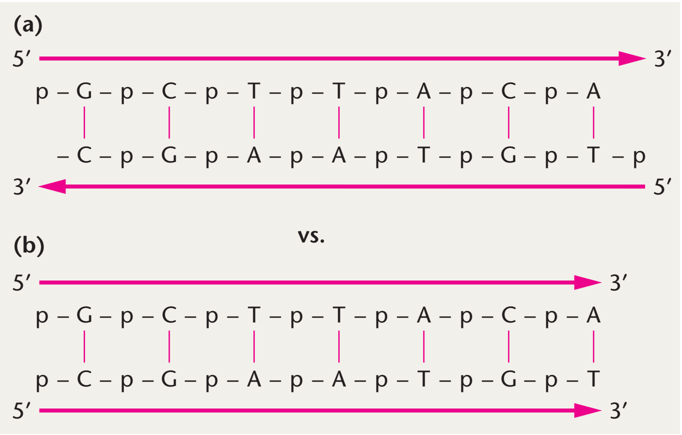


31. To gauge the fidelity of DNA synthesis, Arthur Kornberg and colleagues devised a technique called nearest-neighbor analysis, which determines the frequency with which any two bases occur adjacent to each other along the polynucleotide chain (*J. Biol. Chem.* 236: 864–875). This test relies on the enzyme spleen phosphodiesterase (see the previous problem). As we saw in Figure 11-8, DNA is synthesized by polymerization of 5′-nucleotides—that is, each nucleotide is added with the phosphate on the C-5′ of deoxyribose. However, as shown in the accompanying figure, the phosphodiesterase enzyme cleaves DNA between the phosphate and the C-5′ atom, thereby producing 3′-nucleotides. In this test, the phosphates on only one of the four nucleotide precursors of DNA (cytidylic acid, for example) are made radioactive with ^{32}P, and DNA is synthesized. Then the DNA is subjected to enzymatic cleavage, in which the radioactive phosphate is transferred to the base that is the "nearest neighbor" on the 5′ side of all cytidylic acid nucleotides.

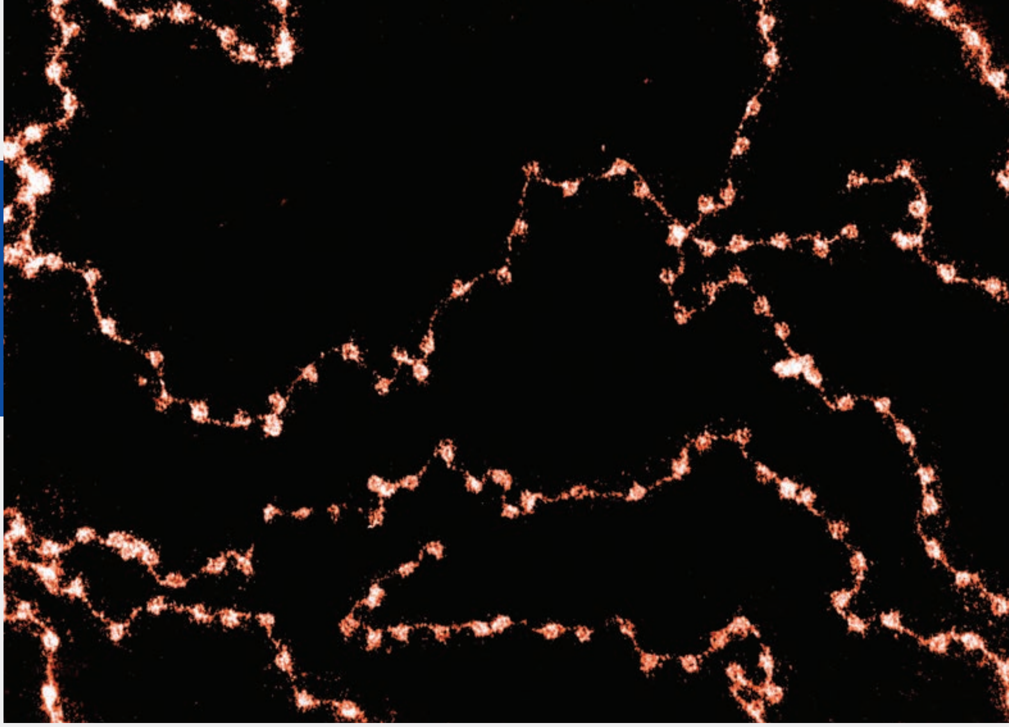

P
O
T
^{32}P
C
G
DNA synthesis
Nucleotides prior to synthesis; cytidylic acid nucleotides are labeled with ^{32}P
Enzyme treatment
Cleavage between phosphates and 5′-carbons
^{32}p transferred to C's nearest neighbors (T and G)

Following four separate experiments, in each of which a different one of the four nucleotide types is radioactive, the frequency of all 16 possible nearest neighbors can be calculated. When Kornberg applied the nearest-neighbor frequency test to the DNA template and resultant product from a variety of experiments, he found general agreement between the nearest-neighbor frequencies of the two.

Analysis of nearest-neighbor data led Kornberg to conclude that the two strands of the double helix are in opposite polarity to one another. Demonstrate this approach by determining the outcome of such an analysis if the strands of DNA shown here are (a) antiparallel versus (b) parallel:


(a)
5′ 3′
p – G – p – C – p – T – p – T – p – A – p – C – p – A
– C – p – G – p – A – p – A – p – T – p – G – p – T – p
3′ 5′
vs.
(b)
5′ 3′
p – G – p – C – p – T – p – T – p – A – p – C – p – A
p – C – p – G – p – A – p – A – p – T – p – G – p – T
5′ 3′

12



DNA Organization in Chromosomes

Chromatin fibers that have spilled out of a chicken erythrocyte, viewed using electron microscopy.

CHAPTER CONCEPTS

- Genetic information in viruses, bacteria, mitochondria, and chloroplasts, with some exceptions, is contained in a short, circular DNA molecule relatively free of associated proteins.
- Eukaryotic cells, in contrast to viruses and bacteria, contain large amounts of DNA that during most of the cell cycle is organized into nucleosomes and is present as either uncoiled chromatin fibers or more condensed structures.
- The uncoiled chromatin fibers characteristic of interphase coil up and condense into chromosomes during the stages of eukaryotic cell division.
- Whereas bacterial genomes consist of mostly unique DNA sequences coding for proteins, eukaryotic genomes contain a mixture of both unique and repetitive DNA sequences.
- Eukaryotic genomes consist mostly of noncoding DNA sequences.

Once geneticists understood that DNA houses genetic information, they focused their energies on discovering how DNA is organized into genes and how these basic units of genetic function are organized into chromosomes. In short, the next major questions they tackled had to do with how the genetic material is organized within the genome. These issues have a bearing on many areas of genetic inquiry. For example, the ways in which genomic organization varies has a major bearing on the regulation of genetic expression.

In this chapter, we focus on the various ways DNA is organized into chromatin, which in turn is organized into chromosomes. These structures have been studied using numerous approaches, including biochemical analysis as well as visualization by light microscopy and electron microscopy. In the first part of the chapter, after surveying what we know about chromosomes in viruses and bacteria, we examine two specialized eukaryotic structures called polytene and lampbrush chromosomes. Then, we turn to a consideration of how eukaryotic chromosomes are organized at the molecular level—for example, how DNA is complexed with proteins to form chromatin and how the chromatin fibers characteristic of interphase are condensed into chromosome structures visible during mitosis and meiosis. We conclude the chapter by examining certain aspects of DNA sequence organization in eukaryotic genomes.

12.1 Viral and Bacterial Chromosomes Are Relatively Simple DNA Molecules

The chromosomes of viruses and bacteria are much less complicated than those in eukaryotes. They usually consist of a single nucleic acid molecule quite different from the multiple chromosomes constituting the genome of higher forms. Bacterial chromosomes are largely devoid of associated proteins and contain relatively less genetic information. These characteristics have greatly simplified genetic analysis in these organisms, and we now have a fairly comprehensive view of the structure of their chromosomes.

The chromosomes of viruses consist of a nucleic acid molecule—either DNA or RNA—that can be either single or double stranded. They can exist as circular structures (covalently closed circles), or they can take the form of linear molecules. The single-stranded DNA of the **ϕX174 bacteriophage** and the double-stranded DNA of the polyoma virus are closed circles housed within the protein coat of the mature viruses. The **bacteriophage lambda (λ),** on the other hand, possesses a linear double-stranded DNA molecule prior to infection, but it closes to form a ring upon infection of the host cell. Still other viruses, such as the T-even series of bacteriophages, have linear double-stranded chromosomes of DNA that do not form circles inside the bacterial host. Thus, circularity is not an absolute requirement for replication in viruses.

Viral nucleic acid molecules have been visualized with the electron microscope. **Figure 12.1** shows a mature bacteriophage λ and its double-stranded DNA molecule in the circular configuration. One constant feature shared by viruses, bacteria, and eukaryotic cells is the ability to package an exceedingly long DNA molecule into a relatively small volume. In λ, the DNA is 17 μm long and must fit into the phage head, which is less than 0.1 μm on any side. **Table 12.1** compares the length of the chromosomes of several viruses with the size of their head structure. In each case, a similar packaging feat must be accomplished. Compare the dimensions given for phage T2 with the micrograph of both the DNA and the viral particle shown in **Figure 12.2**. Seldom does the space available in the head of a virus exceed the chromosome volume by more than a factor of two. In many cases, almost all space is filled, indicating nearly perfect packing. Once packed within the head, the virus's genetic

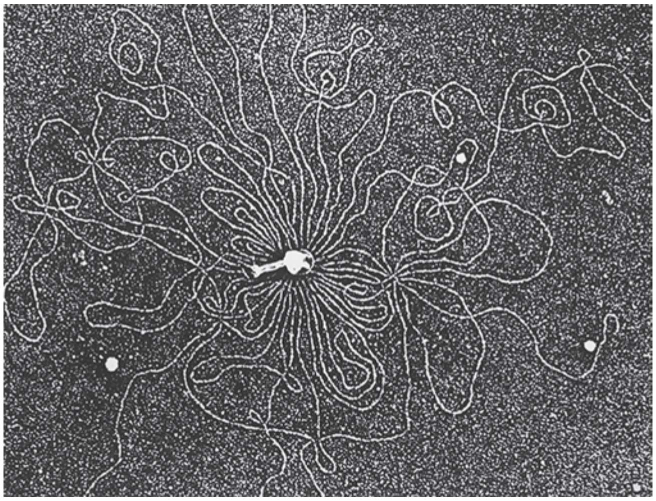


FIGURE 12.2 Electron micrograph of bacteriophage T2, which has had its DNA released by osmotic shock. The chromosome is 52 μm long.

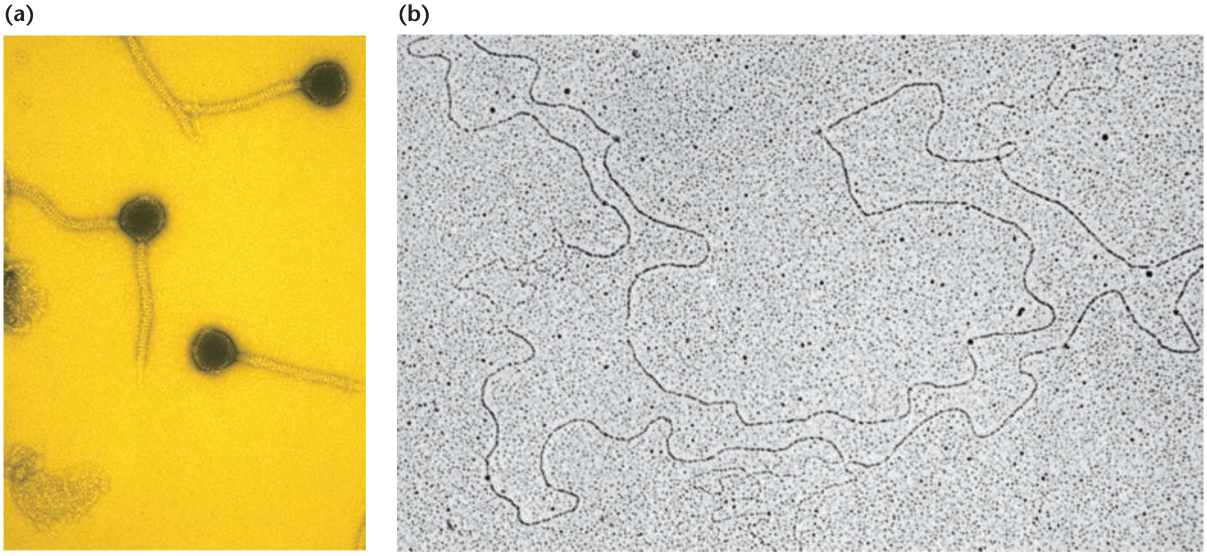


FIGURE 12.1 Electron micrographs of (a) phage λ and (b) the DNA isolated from it. The chromosome is 17 μm long. The phages are magnified about five times more than the DNA.

TABLE 12.1 The Genetic Material of Representative Viruses and Bacteria

Source		Nucleic Acid			Overall Size of Viral Head or Bacterial Cell (μm)
		Type	SS or DS*	Length (μm)	
Viruses	ΦX174	DNA	SS	2.0	0.025 × 0.025
	Tobacco mosaic virus	RNA	SS	3.3	0.30 × 0.02
	Lambda phage	DNA	DS	17.0	0.07 × 0.07
	T2 phage	DNA	DS	52.0	0.07 × 0.10
Bacteria	*Haemophilus influenzae*	DNA	DS	832.0	1.00 × 0.30
	Escherichia coli	DNA	DS	1200.0	2.00 × 0.50

*SS = single-stranded; DS = double-stranded

material is functionally inert until it is released into a host cell.

Bacterial chromosomes are also relatively simple in form. They always consist of a double-stranded DNA molecule, compacted into a structure sometimes referred to as the **nucleoid.** *Escherichia coli,* the most extensively studied bacterium, has a large, circular chromosome measuring approximately 1200 μm (1.2 mm) in length. When the cell is gently lysed and the chromosome is released, the chromosome can be visualized under the electron microscope (Figure 12.3).

DNA in bacterial chromosomes is found to be associated with several types of DNA-binding proteins. Two, called **HU** and **H-NS (histone-like nucleoid structuring) proteins,** are small but abundant in the cell and contain a high percentage of positively charged amino acids that can bond ionically to the negative charges of the phosphate groups in DNA. These proteins function to fold and bend DNA. As such, coils are created that have the effect of compacting the DNA constituting the nucleoid. Additionally, H-NS proteins, like histones in eukaryotes, have been implicated in regulating gene activity in a nonspecific way.

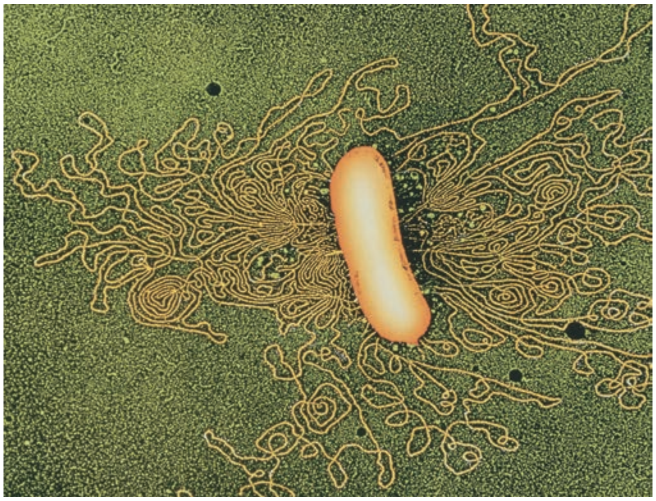


FIGURE 12.3 Electron micrograph of the bacterium *Escherichia coli,* which has had its DNA released by osmotic shock. The chromosome is 1200 μm long.

NOW SOLVE THIS

12.1 In bacteriophages and bacteria, the DNA is almost always organized into circular (closed loops) chromosomes. Phage λ is an exception, maintaining its DNA in a linear chromosome within the viral particle. However, as soon as this DNA is injected into a host cell, it circularizes before replication begins. What advantage exists in replicating circular DNA molecules compared to linear molecules, characteristic of eukaryotic chromosomes?

■ **HINT:** *This problem involves an understanding of eukaryotic DNA replication, as discussed earlier in the text (see Chapter 11). The key to its solution is to consider why the enzyme telomerase is essential in eukaryotic DNA replication, and why bacterial and viral chromosomes can be replicated without encountering the "telomere" problem.*

12.2 Supercoiling Facilitates Compaction of the DNA of Viral and Bacterial Chromosomes

One major insight into the way DNA is organized and packaged in viral and bacterial chromosomes has come from the discovery of **supercoiled DNA,** which is particularly characteristic of closed-circular molecules. Supercoiled DNA was first proposed as a result of a study of double-stranded DNA molecules derived from the polyoma virus, which causes tumors in mice. In 1963, it was observed that when the polyoma DNA was subjected to high-speed centrifugation, it was resolved into three distinct components, each of different density and compactness. The one that was least dense and thus least compact was subsequently proposed by Jerome Vinograd to consist of linear DNA molecules. The other two fractions were more dense, and both are now known to consist of circular DNA molecules. Closed-circular molecules are more compact and sediment more rapidly during centrifugation than do the same molecules in linear form. All three were of identical molecular weight.

Vinograd proposed further that the denser of the two fractions of circular molecules consisted of covalently closed

DNA helices that are slightly *underwound* in comparison to the less dense circular molecules. Energetic forces stabilizing the double helix resist this underwinding, causing the molecule as a whole to **supercoil,** that is, to contort in a certain way, in order to retain normal base pairing. Vinograd proposed that it is the supercoiled shape that causes tighter packing and thus the increase in density.

The transitions just described are illustrated in Figure 12.4. Consider a double-stranded linear molecule existing in the normal Watson—Crick right-handed helix [Figure 12.4(a)]. This helix contains 20 complete turns, which means the **linking number (*L*)** of this molecule is 20, or $L = 20$. Suppose we were to change this linear molecule into a closed circle by bringing the opposite ends together and joining them [Figure 12.4(b)]. If the closed circle still has a linking number of 20 (if we haven't introduced or eliminated any turns when joining the ends), we define the molecule as being *energetically relaxed.* Now suppose that the circle were subsequently cut open, underwound by two full turns, and then resealed [Figure 12.4(c)]. Such a structure, in which L has been changed to 18, would be *energetically strained* and, as a result, would change its form to relieve the strain.

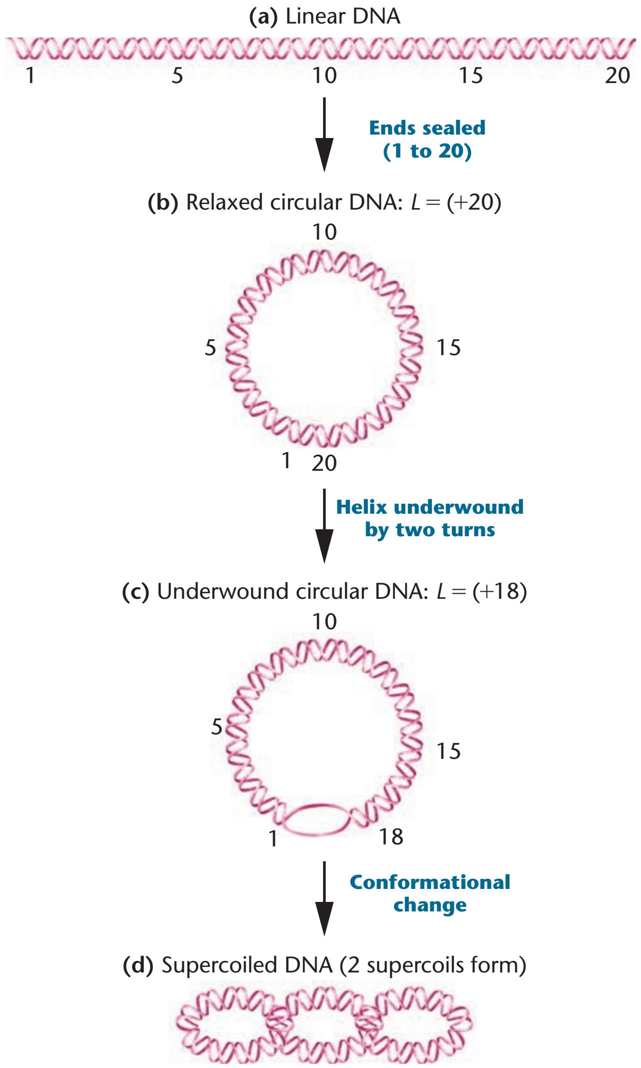

(a) Linear DNA
1
5
10
15
20
Ends sealed (1 to 20)
(b) Relaxed circular DNA: L = (+20)
10
5
15
1
20
Helix underwound by two turns
(c) Underwound circular DNA: L = (+18)
10
5
15
1
18
Conformational change
(d) Supercoiled DNA (2 supercoils form)

FIGURE 12.4 Depictions of the transformations leading to the supercoiling of circular DNA. L signifies the linking number.

In order to assume a more energetically favorable conformation, an underwound molecule will form supercoils in the direction opposite to that of the underwinding. In our case [Figure 12.4(d)], two negative supercoils are introduced spontaneously, reestablishing, in total, the original number of turns in the helix. Use of the term *negative* refers to the fact that, by definition, the supercoils are left-handed (whereas the helix is right-handed). The end result is the formation of a more compact structure with enhanced physical stability.

In most closed-circular DNA molecules in bacteria and their phages, the DNA helix is slightly underwound [as in Figure 12.4(c)]. For example, the virus SV40 contains 5200 base pairs. In energetically relaxed DNA, 10.4 base pairs occupy each complete turn of the helix, and the linking number can be calculated as

$$L = \frac{5200}{10.4} = 500$$

However, analysis of circular SV40 DNA reveals that it is underwound by 25 turns, so L is equal to only 475. Predictably, 25 negative supercoils are observed in the molecule. In *E. coli,* an even larger number of supercoils is observed, greatly facilitating chromosome condensation in the nucleoid region.

Two otherwise identical molecules that differ only in their linking number are said to be **topoisomers** of one another. But how can a molecule convert from one topoisomer to the other if there are no free ends, as is the case in closed circles of DNA? Biologically, this may be accomplished by any one of a group of enzymes that cut one or both of the strands and wind or unwind the helix before resealing the ends.

Appropriately, these enzymes are called **topoisomerases.** First discovered by Martin Gellert and James Wang, these catalytic molecules are known as either type I or type II, depending on whether they cleave one or both strands in the helix, respectively. In *E. coli,* topoisomerase I serves to reduce the number of negative supercoils in a closed-circular DNA molecule. Topoisomerase II introduces negative supercoils into DNA. This latter enzyme is thought to bind to DNA, twist it, cleave both strands, and then pass them through the loop that it has created. Once the phosphodiester bonds are reformed, the linking number is decreased and one or more supercoils form spontaneously.

Supercoiled DNA and topoisomerases are also found in eukaryotes. While the chromosomes in these organisms are not usually circular, supercoils can occur when areas of

DNA are embedded in a lattice of proteins associated with the chromatin fibers. This association creates "anchored" ends, providing the stability for the maintenance of supercoils once they are introduced by topoisomerases. In both prokaryotes and eukaryotes, DNA replication and transcription create supercoils downstream as the double helix unwinds and becomes accessible to the appropriate enzyme.

Topoisomerases may play still other genetic roles involving eukaryotic DNA conformational changes. Interestingly, these enzymes are involved in separating (decatenating) the DNA of sister chromatids following replication.

12.3 Specialized Chromosomes Reveal Variations in the Organization of DNA

We now consider two cases of genetic organization that demonstrate specialized forms that eukaryotic chromosomes can take. Both types—*polytene chromosomes* and *lampbrush chromosomes*—are so large that their organization was discerned using light microscopy long before we understood how mitotic chromosomes form from interphase chromatin. The study of these chromosomes provided many of our initial insights into the arrangement and function of the genetic information. It is important to know that polytene and lampbrush chromosomes are unusual and not typically found in most eukaryotic cells, but the study of their structure has revealed many common themes of chromosome organization.

Polytene Chromosomes

Giant **polytene chromosomes** are found in various tissues (salivary, midgut, rectal, and malpighian excretory tubules) in the larvae of some flies, as well as in several species of protozoans and plants. They were first observed by E. G. Balbiani in 1881. The large amount of information obtained from studies of these genetic structures provided a model system for subsequent investigations of chromosomes. What is particularly intriguing about polytene chromosomes is that they can be seen in the nuclei of interphase cells.

Each polytene chromosome is 200–600 μm long, and when they are observed under the light microscope, they exhibit a linear series of alternating bands and interbands (Figures 12.5 and 12.6). The banding pattern is distinctive for each chromosome in any given species. Individual bands are sometimes called **chromomeres,** a more generalized term describing lateral condensations of material along the axis of a chromosome.

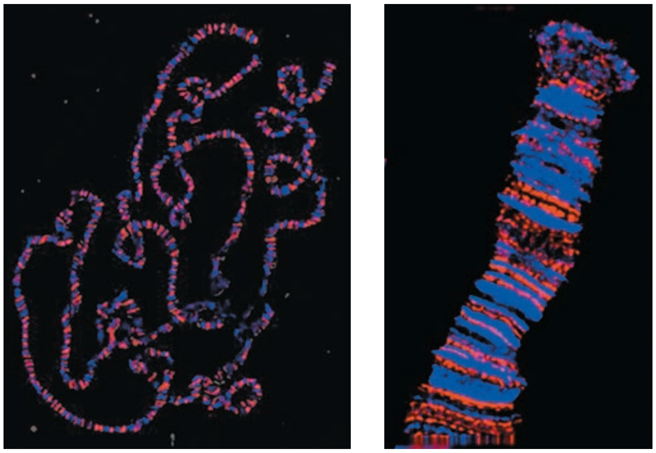


FIGURE 12.5 Polytene chromosomes derived from larval salivary gland cells of *Drosophila*.

Extensive study using electron microscopy and radioactive tracers led to an explanation for the unusual appearance of these chromosomes. First, polytene chromosomes represent paired homologs. This in itself is highly unusual, since they are present in somatic cells, where, in most organisms, chromosomal material is normally dispersed as chromatin and homologs are not paired. Second, their large size and distinctive appearance result from their being composed of large numbers of identical DNA strands. The DNA of these paired homologs undergoes many rounds of replication, *but without strand separation or cytoplasmic division.* As replication proceeds, chromosomes are created, having 1000–5000 DNA strands that remain in precise parallel alignment with one another. Apparently, the parallel register of so many DNA strands gives rise to the distinctive band pattern along the axis of the chromosome.

The presence of bands on polytene chromosomes was initially interpreted as the visible manifestation of individual genes. The discovery that the strands present in bands undergo localized uncoiling during genetic activity further

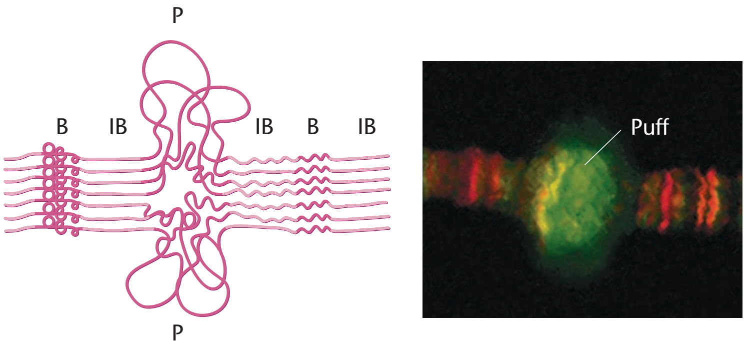


FIGURE 12.6 Photograph of a puff within a polytene chromosome. The diagram depicts the uncoiling of strands within a band region (B) to produce a puff (P) in a polytene chromosome. Each band (B) represents a chromomere. Interband regions (IBs) are also labeled.

strengthened this view. Each such uncoiling event results in a bulge called a **puff,** so named because of its appearance under the microscope (Figure 12.6). That puffs are visible manifestations of a high level of gene activity (transcription that produces RNA) is evidenced by their high rate of incorporation of radioactively labeled RNA precursors, as assayed by autoradiography. Bands that are not extended into puffs incorporate fewer radioactive precursors or none at all.

The study of bands during development in insects such as *Drosophila* and the midge fly *Chironomus* reveals differential gene activity. A characteristic pattern of band formation that is equated with gene activation is observed as development proceeds. Despite attempts to resolve the issue, it is not yet clear how many genes are contained in each band. However, we do know that in *Drosophila,* which contains about 15,000 genes, there are approximately 5000 bands. Interestingly, a band may contain up to 10^7 base pairs of DNA, enough DNA to encode 50 to 100 average-size genes.

NOW SOLVE THIS

12.2 After salivary gland cells from *Drosophila* are isolated and cultured in the presence of radioactive thymidylic acid, autoradiography is performed, revealing polytene chromosomes. Predict the distribution of the grains along the chromosomes.

HINT: *This problem involves an understanding of the organization of DNA in polytene chromosomes. The key to its solution is to be aware that ^{3}H-thymidine, as a molecular tracer, will only be incorporated into DNA during its replication.*

Lampbrush Chromosomes

Another specialized chromosome that has given us insight into chromosomal structure is the **lampbrush chromosome,** so named because it resembles the brushes used to clean kerosene lamp chimneys in the nineteenth century. Lampbrush chromosomes were first discovered in 1882 by Walther Flemming in salamander oocytes and then seen in 1892 by J. Ruckert in shark oocytes. They are now known to be characteristic of most vertebrate oocytes as well as the spermatocytes of some insects. Therefore, they are meiotic chromosomes. Most of the experimental work on them has been done with material taken from amphibian oocytes.

These unique chromosomes are easily isolated from oocytes in the first prophase of meiosis, where they are active in directing the metabolic activities of the developing cell. The homologs are seen as synapsed pairs held together by chiasmata. However, instead of condensing, as most meiotic chromosomes do, lampbrush chromosomes are often extended to lengths of 500 to 800 μm. Later, in meiosis, they revert to their normal length of 15 to 20 μm. Based on these observations, lampbrush chromosomes are interpreted as being extended, uncoiled versions of the normal meiotic chromosomes.

The two views of lampbrush chromosomes in Figure 12.7 provide significant insights into their morphology. Part (a) shows the meiotic configuration under the light microscope. The linear axis of each horizontal structure seen in the figure contains a large number of condensed areas, which, as with polytene chromosomes, are referred to as *chromomeres.* Emanating from each chromomere is a pair of lateral loops, giving the chromosome its distinctive appearance. In part (b), the scanning electron micrograph (SEM) shows many adjacent pairs of loops in detail along one of the axes. As with bands in polytene chromosomes, much more DNA is present in each loop than is needed to encode a single gene. Such an SEM provides a clear view of the chromomeres and the chromosomal fibers emanating from them. Each chromosomal loop is thought to be composed of one DNA double helix, while the central

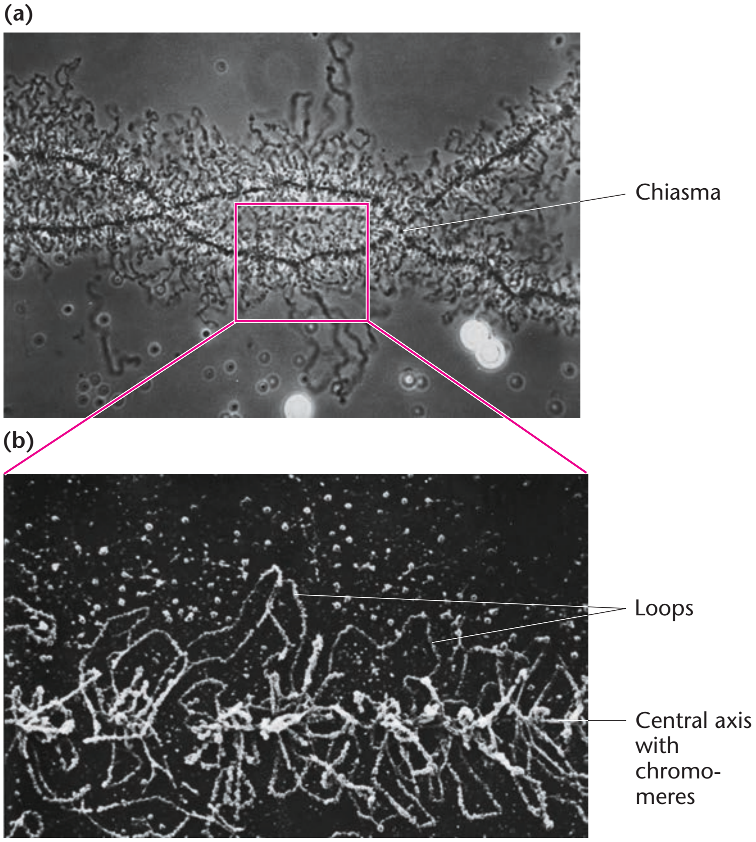


FIGURE 12.7 Lampbrush chromosomes derived from amphibian oocytes. (a) A photomicrograph. (b) A scanning electron micrograph.

axis is made up of two DNA helices. This hypothesis is consistent with the belief that each meiotic chromosome is composed of a pair of sister chromatids. Studies using radioactive RNA precursors reveal that the loops are active in the synthesis of RNA. The lampbrush loops, in a manner similar to puffs in polytene chromosomes, represent DNA that has been reeled out from the central chromomere axis during transcription.

12.4 DNA Is Organized into Chromatin in Eukaryotes

We now turn our attention to the way DNA is organized in eukaryotic chromosomes, which are most clearly visible as highly condensed structures during mitosis. However, after chromosome separation and cell division, cells enter the interphase stage of the cell cycle, at which time the components of the chromosome uncoil and decondense into a form referred to as **chromatin.** While in interphase, the chromatin is dispersed throughout the nucleus. As the cell cycle progresses, cells may replicate their DNA and reenter mitosis, whereupon chromatin coils and condenses back into visible chromosomes once again. This condensation represents a length contraction of some 10,000 times for each chromatin fiber.

The organization of DNA during the transitions just described is much more intricate and complex than in viruses or bacteria, which never exhibit a complex process similar to mitosis. This is due to the greater amount of DNA per chromosome in eukaryotes, as well as the presence of a large number of proteins associated with eukaryotic DNA. For example, while DNA in the *E. coli* chromosome is 1200 μm long, the DNA in each human chromosome ranges from 19,000 to 73,000 μm in length. In a single human nucleus, all 46 chromosomes contain sufficient DNA to extend almost 2 meters. This genetic material, along with its associated proteins, is contained within a nucleus that usually measures about 5 to 10 μm in diameter.

Chromatin Structure and Nucleosomes

As we have seen, the genetic material of viruses and bacteria consists of strands of DNA or RNA relatively devoid of proteins. In contrast, eukaryotic chromatin has a substantial amount of protein associated with the chromosomal DNA in all phases of the cell cycle. The associated proteins can be categorized as either positively charged **histones** or less positively charged *nonhistone proteins.* Of these two groups, the histones play the most essential structural role. Histones contain large amounts of the positively charged amino acids lysine and arginine, making it possible for them to bond electrostatically to the negatively charged phosphate groups of nucleotides. Recall that a similar interaction has been proposed for several bacterial proteins. The five main types of histones are shown in **Table 12.2**.

The general model for chromatin structure is based on the assumption that chromatin fibers, composed of DNA and protein, undergo extensive coiling and folding as they are condensed within the cell nucleus. Moreover, X-ray diffraction studies confirm that histones play an important role in chromatin structure. Chromatin produces regularly spaced diffraction rings, suggesting that repeating structural units occur along the chromatin axis. If the histone molecules are chemically removed from chromatin, the regularity of this diffraction pattern is disrupted.

A basic model for chromatin structure was worked out in the mid-1970s. The following observations were particularly relevant to the development of this model:

1. Digestion of chromatin by certain endonucleases, such as micrococcal nuclease, yields DNA fragments that are approximately 200 base pairs in length or multiples thereof. This enzymatic digestion is not random, for if it were, we would expect a wide range of fragment sizes. Thus, chromatin consists of some type of repeating unit, each of which protects the DNA from enzymatic cleavage except where any two units are joined. It is the area between units that is attacked and cleaved by the endonuclease.
2. Electron microscopic observations of chromatin have revealed that chromatin fibers are composed of linear arrays of spherical particles (**Figure 12.8**). Discovered by Ada and Donald Olins, the particles occur regularly along the axis of a chromatin strand and resemble beads on a string. These particles, initially referred to as *v-bodies* (*v* is the Greek letter nu), are now called **nucleosomes.** These findings conform to the above observation that suggests the existence of repeating units.
3. Studies of the chemical association between histone molecules and DNA in the nucleosomes of chromatin show that histones H2A, H2B, H3, and H4 occur as two types of tetramers, $(H2A)_2 \cdot (H2B)_2$ and $(H3)_2 \cdot (H4)_2$. Roger Kornberg predicted that each repeating nucleosome unit consists of one of each tetramer (creating an octomer) in association with about 200 base pairs of DNA. Such a

TABLE 12.2 Categories and Properties of Histone Proteins

Histone Type	Lysine–Arginine Content	Molecular Weight (Da)
H1	Lysine-rich	23,000
H2A	Slightly lysine-rich	14,000
H2B	Slightly lysine-rich	13,800
H3	Arginine-rich	15,300
H4	Arginine-rich	11,300

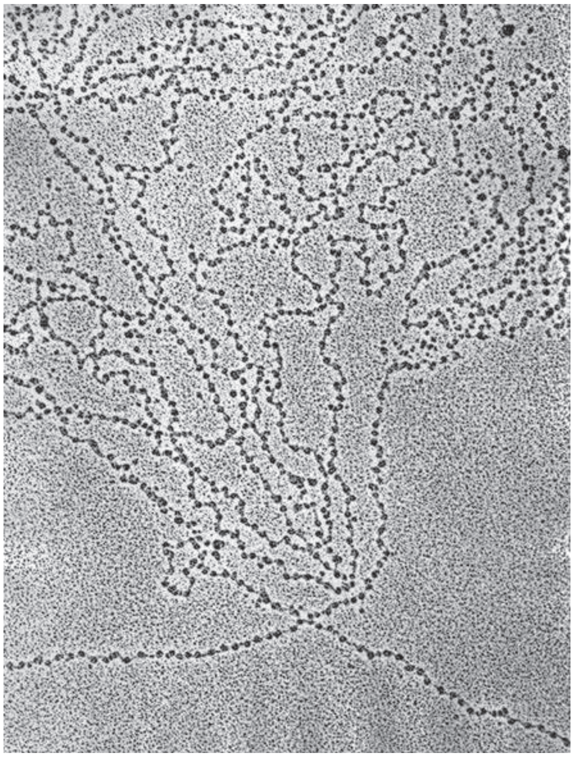


FIGURE 12.8 An electron micrograph revealing nucleosomes appearing as "beads on a string" along chromatin strands derived from *Drosophila melanogaster*.

structure is consistent with previous observations and provides the basis for a model that explains the interaction of histones and DNA in chromatin.

4. When nuclease digestion time is extended, some of the 200 base pairs of DNA are removed from the nucleosome, creating what is called a **nucleosome core particle** consisting of 147 base pairs. The DNA lost in the prolonged digestion is responsible for linking nucleosomes together. This **linker DNA** is associated with the fifth histone, H1.

On the basis of this information, as well as on X-ray and neutron-scattering analyses of crystallized core particles by John T. Finch, Aaron Klug, and others, a detailed model of the nucleosome was put forward in 1984, providing a basis for predicting chromatin structure and its condensation into chromosomes. In this model, illustrated in Figure 12.9, a 147-bp length of the 2-nm-diameter DNA molecule coils around an octamer of histones in a left-handed superhelix that completes about 1.7 turns per nucleosome. Each nucleosome, ellipsoidal in shape, measures about 11 nm at its longest point [Figure 12.9(a)]. Significantly, the formation of the nucleosome represents the first level of packing, whereby the DNA helix is reduced to about one-third of its original length by winding around the histones.

In the nucleus, the chromatin fiber seldom, if ever, exists in the extended form described in the previous paragraph (that is, as an extended chain of nucleosomes). Instead, the 11-nm-diameter fiber is further packed into a thicker structure, initially called a *solenoid*, but now referred to as a *30-nm fiber*. [Figure 12.9(b)]. This thicker structure, which is dependent on the presence of histone H1, consists of numerous nucleosomes coiled around and stacked upon one another, creating a second level of packing. This provides a sixfold increase in compaction of the DNA. It is this structure that is characteristic of an uncoiled chromatin fiber in interphase of the cell cycle. In the transition to the mitotic chromosome, still further compaction must occur. The 30-nm structures are folded into a series of *looped domains,* which further condense the chromatin fiber into a structure that is 300 nm in diameter [Figure 12.9(c)]. These *coiled chromatin fibers* are then compacted into the chromosome arms that constitute a chromatid, one of the longitudinal subunits of the metaphase chromosome [Figure 12.9(d)]. While Figure 12.9 shows the chromatid arms to be 700 nm in diameter, this value undoubtedly varies among different organisms. At a value of 700 nm, a pair of sister chromatids comprising a chromosome measures about 1400 nm.

The importance of the organization of DNA into chromatin and of chromatin into mitotic chromosomes can be illustrated by considering that a human cell stores its genetic material in a nucleus about 5 to 10 μm in diameter. The haploid genome contains more than 3 billion base pairs of DNA distributed among 23 chromosomes. The diploid cell contains twice that amount. At 0.34 nm per base pair, this amounts to an enormous length of DNA (as stated earlier, almost 2 meters)! One estimate is that the DNA inside a typical human nucleus is complexed with roughly 25×10^6 nucleosomes.

In the overall transition from a fully extended DNA helix to the extremely condensed status of the mitotic chromosome, a packing ratio (the ratio of DNA length to the length of the structure containing it) of about 500 to 1 must be achieved. In fact, our model accounts for a ratio of only about 50 to 1. Obviously, the larger fiber can be further bent, coiled, and packed to achieve even greater condensation during the formation of a mitotic chromosome.

NOW SOLVE THIS

12.3 If a human nucleus is 10 μm in diameter, and it must hold as much as 2 m of DNA, which is complexed into nucleosomes that during full extension are 11 nm in diameter, what percentage of the volume of the nucleus is occupied by the genetic material?

■ **HINT:** *This problem asks you to make some numerical calculations in order to see just how "filled" the eukaryotic nucleus is with a diploid amount of DNA. The key to its solution is the use of the formula* $V = (4/3)\pi r^3$, *which calculates the volume of a sphere.*

For more practice, see Problems 12–14.

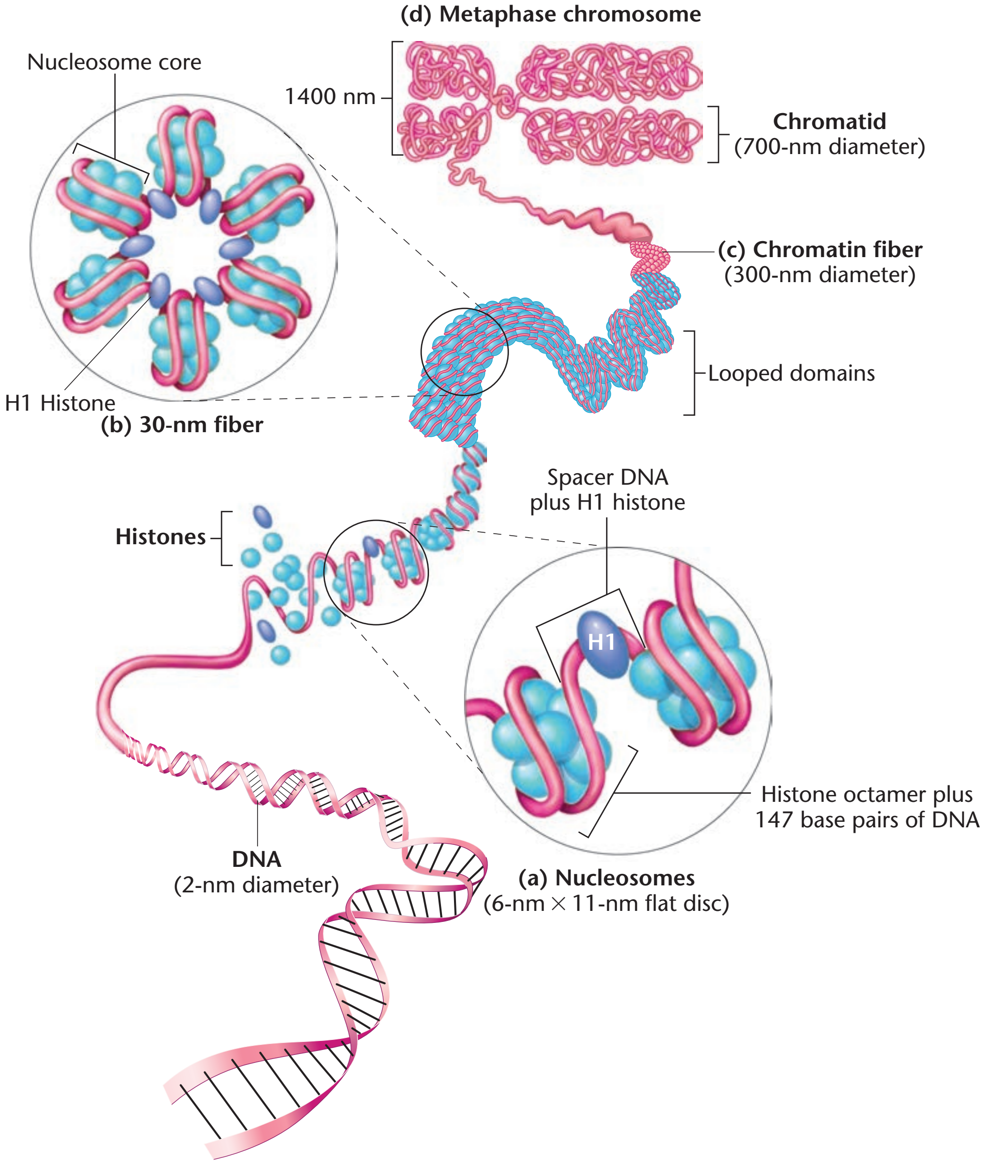


FIGURE 12.9 General model of the association of histones and DNA to form nucleosomes, illustrating the way in which each thickness of fiber may be coiled into a more condensed structure, ultimately producing a metaphase chromosome.

Chromatin Remodeling

As with many significant endeavors in genetics, the study of nucleosomes has answered some important questions but at the same time has opened up new ones. For example, in the preceding discussion, we established that histone proteins play an important structural role in packaging DNA into the nucleosomes that make up chromatin. While this discovery helped solve the structural problem of how the huge amount of DNA is organized within the eukaryotic nucleus, it brought another problem to the fore: *When present in several levels of compaction within the chromatin fiber, DNA is inaccessible to interaction with other important DNA-binding proteins.* For example, the various proteins that function in enzymatic and regulatory roles during the processes of replication and transcription must interact directly with DNA. To accommodate these protein—DNA interactions, chromatin must be induced to change its structure, a process now referred to as **chromatin remodeling.** To allow replication and gene expression, chromatin must relax its compact structure and expose regions of DNA to these proteins, and there must also be a mechanism for reversing the process during periods of inactivity.

Insights into how different states of chromatin structure might be achieved began to emerge in 1997, when Timothy Richmond and members of his research team were able to significantly improve the level of resolution in X-ray diffraction studies of nucleosome crystals, from 7 Å in the 1984 studies to 2.8 Å in the 1997 studies. One model based on their work is shown in Figure 12.10. At this resolution,

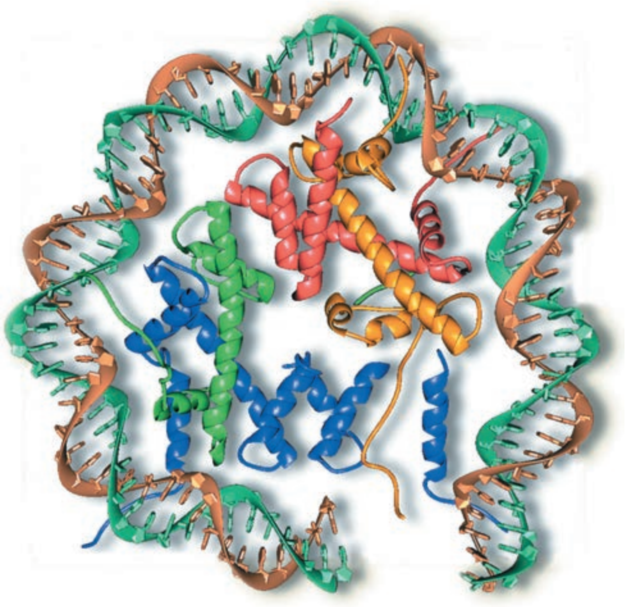


FIGURE 12.10 The nucleosome core particle derived from X-ray crystal analysis at 2.8 Å resolution. The double-helical DNA surrounds four pairs of histones.

most atoms are visible, thus revealing the subtle twists and turns of the superhelix of DNA encircling the histones. Recall that the double-helical ribbon in the figure represents 147 bp of DNA surrounding four pairs of histone proteins. This configuration is essentially repeated over and over in the chromatin fiber and is the principal packaging unit of DNA in the eukaryotic nucleus.

By 2003, Richmond and colleagues achieved a resolution of 1.9 Å that revealed the details of the location of each histone entity within the nucleosome. Of particular relevance to the discussion of chromatin remodeling is the observation that there are unstructured **histone tails** that are *not* packed into the folded histone domains within the core of the nucleosomes but instead protrude from it. For example, tails devoid of any secondary structure extending from histones H3 and H2B protrude through the minor-groove channels of the DNA helix. You should look carefully at Figure 12.10 and locate examples of such tails. Other tails of histone H4 appear to make a connection with adjacent nucleosomes. The significance of histone tails is that they provide potential targets along the chromatin fiber for a variety of chemical modifications that may be linked to genetic functions, including chromatin remodeling and the possible regulation of gene expression.

Several of these potential chemical modifications are now recognized as important to genetic function. One of the most well-studied histone modifications involves **acetylation** by the action of the enzyme *histone acetyltransferase (HAT).* The addition of an acetyl group to the positively charged amino group present on the side chain of the amino acid lysine effectively changes the net charge of the protein by neutralizing the positive charge. Lysine is in abundance in histones, and geneticists have known for some time that acetylation is linked to gene activation. It appears that high levels of acetylation open up, or remodel, the chromatin fiber, an effect that increases in regions of active genes and decreases in inactive regions. In the well-studied example of the inactivation of the X chromosome in mammals, forming a Barr body (Chapter 7), histone H4 is known to be greatly underacetylated.

Two other important chemical modifications are the **methylation** and **phosphorylation** of amino acids that are part of histones. These chemical processes result from the action of enzymes called *methyltransferases* and *kinases,* respectively. Methyl groups can be added to both arginine and lysine residues in histones, and this change has been correlated with gene activity. Phosphate groups can be added to the hydroxyl groups of the amino acids serine and histidine, introducing a negative charge on the protein. During the cell cycle, increased phosphorylation, particularly of histone H3, is known to occur at characteristic times. Such chemical modification is believed to be related to the cycle of chromatin unfolding and condensation that occurs during and after DNA replication. It is important to note that the above chemical modifications (acetylation, methylation, and phosphorylation) are all reversible, under the direction of specific enzymes.

Interestingly, while methylation of histones *within nucleosomes* is often positively correlated with gene activity in eukaryotes, methylation of the nitrogenous base cytosine *within polynucleotide chains of DNA,* forming **5-methyl cytosine,** is usually negatively correlated with gene activity. Methylation of cytosine occurs most often when the nucleotide cytidylic acid is next to the nucleotide guanylic acid, forming what is called a **CpG island.** We must conclude, then, that methylation can have a positive or a negative impact on gene activity.

The research described above has extended our knowledge of nucleosomes and chromatin organization and serves here as a general introduction to the concept of chromatin remodeling. A great deal more work must be done, however, to elucidate the specific involvement of chromatin remodeling during genetic processes. In particular, the way in which the modifications are influenced by regulatory molecules within cells will provide important insights into the mechanisms of gene expression. What is clear is that the dynamic forms in which chromatin exists are vitally important to the way that all genetic processes directly involving DNA are executed. We will return to a more detailed discussion of the role of chromatin remodeling when we consider the regulation of eukaryotic gene expression later in the text (see Chapter 17). In addition, chromatin remodeling is an important topic in the discussion of **epigenetics,** the study of modifications of an organism's genetic and phenotypic expression that are *not* attributable to alteration of the DNA sequence making up a gene. This topic is discussed in depth in a future chapter (see Chapter 19).

Heterochromatin

Although we know that the DNA of the eukaryotic chromosome consists of one continuous double-helical fiber along its entire length, we also know that the whole chromosome is not structurally uniform from end to end. In the early part of the twentieth century, it was observed that some parts of the chromosome remain condensed and stain deeply during interphase, while most parts are partially uncoiled and do not stain. In 1928, the terms **euchromatin** and **heterochromatin** were coined to describe the parts of chromosomes that are uncoiled and those that remain condensed, respectively.

Subsequent investigation revealed a number of characteristics that distinguish heterochromatin from euchromatin. Heterochromatic areas are genetically inactive because they either lack genes or contain genes that are repressed. Also, heterochromatin replicates later during the S phase of the cell cycle than does euchromatin. The discovery of heterochromatin provided the first clues that parts of eukaryotic chromosomes do not always encode proteins. For example, one particular heterochromatic region of the chromosome, the *telomere,* is thought to be involved in maintenance of the chromosome's structural integrity, and another region, the *centromere,* is involved in chromosome movement during cell division.

The presence of heterochromatin is unique to and characteristic of the genetic material of eukaryotes. In some cases, whole chromosomes are heterochromatic. A case in point is the mammalian Y chromosome, much of which is genetically inert. And, as we discussed earlier in the text (see Chapter 7), the inactivated X chromosome in mammalian females is condensed into an inert heterochromatic Barr body. In some species, such as mealy bugs, the chromosomes of one entire haploid set are heterochromatic.

When certain heterochromatic areas from one chromosome are translocated to a new site on the same or another nonhomologous chromosome, genetically active areas sometimes become genetically inert if they now lie adjacent to the translocated heterochromatin. As we saw earlier in the text (see Chapter 4), this influence on existing euchromatin is one example of what is more generally referred to as a **position effect.** That is, the position of a gene or group of genes relative to all other genetic material may affect their expression.

12.5 Chromosome Banding Differentiates Regions along the Mitotic Chromosome

Until about 1970, mitotic chromosomes viewed under the light microscope could be distinguished only by their relative sizes and the positions of their centromeres. Unfortunately, even in organisms with a low haploid number, two or more chromosomes are often visually indistinguishable from one another. Since that time, however, cytological procedures were developed that made possible differential staining along the longitudinal axis of mitotic chromosomes. Such methods are referred to as **chromosome-banding techniques,** because the staining patterns resemble the bands of polytene chromosomes.

One of the first chromosome-banding techniques was devised by Mary-Lou Pardue and Joe Gall. They found that if chromosome preparations from mice were heat denatured and then treated with Giemsa stain, a unique staining pattern emerged: Only the centromeric regions of mitotic chromosomes took up the stain! The staining pattern was thus referred to as **C-banding.** Relevant to our immediate discussion, this cytological technique identifies a specific area of the chromosome composed of heterochromatin. A micrograph of the human karyotype treated in this way is shown in Figure 12.11. Mouse chromosomes are all telocentric, thus localizing the stain at the end of each chromosome.

Other chromosome-banding techniques were developed at about the same time. The most useful of these techniques produces a staining pattern differentially along the length of each chromosome. This method, producing **G-bands**

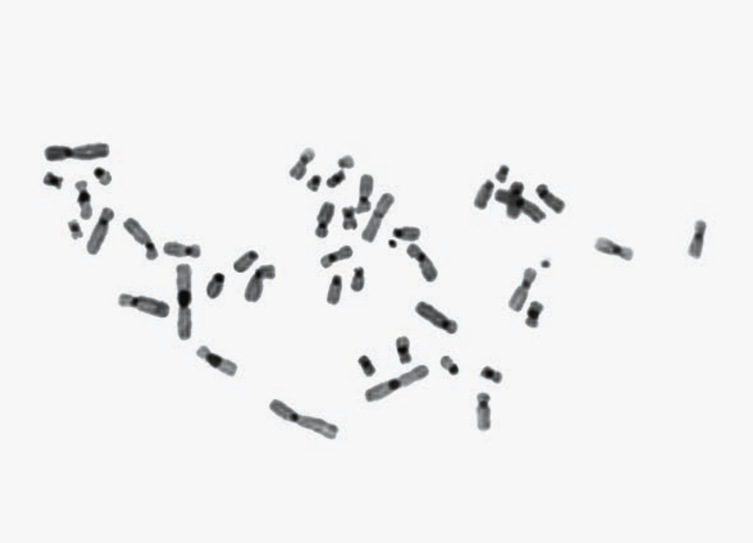


FIGURE 12.11 A human mitotic chromosome preparation processed to demonstrate C-banding. Only the centromeres stain (as small dark circles).

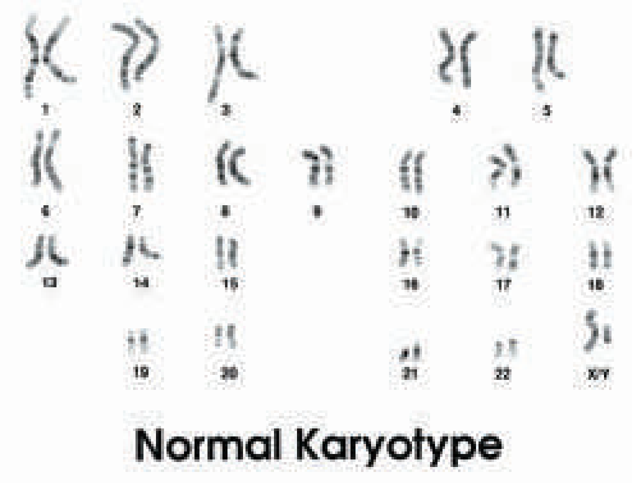


FIGURE 12.12 G-banded karyotype of a normal human male. Chromosomes were derived from cells in metaphase.

(Figure 12.12), involves the digestion of the mitotic chromosomes with the proteolytic enzyme trypsin, followed by Giemsa staining. The differential staining reactions reflect the heterogeneity and complexity of the chromosome along its length.

In 1971 a uniform nomenclature for human chromosome-banding patterns was established based on G-banding. Figure 12.13 illustrates the application of this nomenclature to the X chromosome. On the left of the chromosome are the various organizational levels of banding of the p and q arms that can be identified; the resulting designation for each of the specific regions is shown on the right side.

Although the molecular mechanisms involved in producing the various banding patterns are not well understood, the bands have played an important role in cytogenetic analysis, particularly in humans. The pattern of banding on each chromosome is unique, allowing a distinction to be made even between those chromosomes that are identical in size and centromere placement (e.g., human chromosomes 4 and 5 and 21 and 22). So precise is the banding pattern of each chromosome that homologs can be distinguished from one another, and when a segment of one chromosome has been translocated to another chromosome, its origin can be determined with great precision.

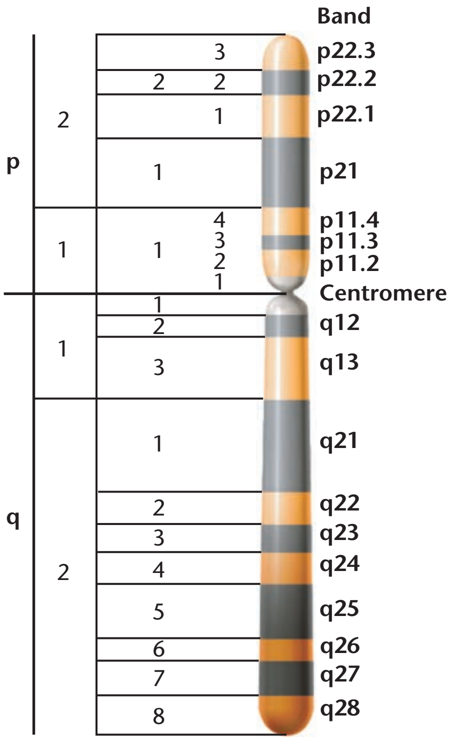


FIGURE 12.13 The regions of the human X chromosome distinguished by its banding pattern. The designations on the right identify specific bands.

12.6 Eukaryotic Genomes Demonstrate Complex Sequence Organization Characterized by Repetitive DNA

Thus far, we have examined the general structure of chromosomes in bacteriophages, bacteria, and eukaryotes. We now begin an examination of what we know about the organization of DNA sequences within the chromosomes making up an organism's genome, placing our emphasis on eukaryotes.

In addition to single copies of unique DNA sequences that make up genes, many DNA sequences within eukaryotic chromosomes are repetitive in nature. Various levels of repetition occur within the genomes of organisms. Many studies have now provided insights into **repetitive DNA,** demonstrating various classes of sequences and organization. Figure 12.14 schematizes these categories. Some functional genes are present in more than one copy (they are referred to as **multiple-copy genes**) and so are repetitive in nature. However, the majority of repetitive sequences do not encode proteins. Nevertheless, many are transcribed, and the resultant RNAs play multiple roles in eukaryotes, including chromatin remodeling, as discussed in detail in a future chapter (see Chapter 18). We will explore three main categories of repetitive sequences: (1) heterochromatin found to be associated with centromeres and making up telomeres, (2) tandem repeats of both short and long DNA sequences, and (3) transposable sequences that are interspersed throughout the genome of eukaryotes.

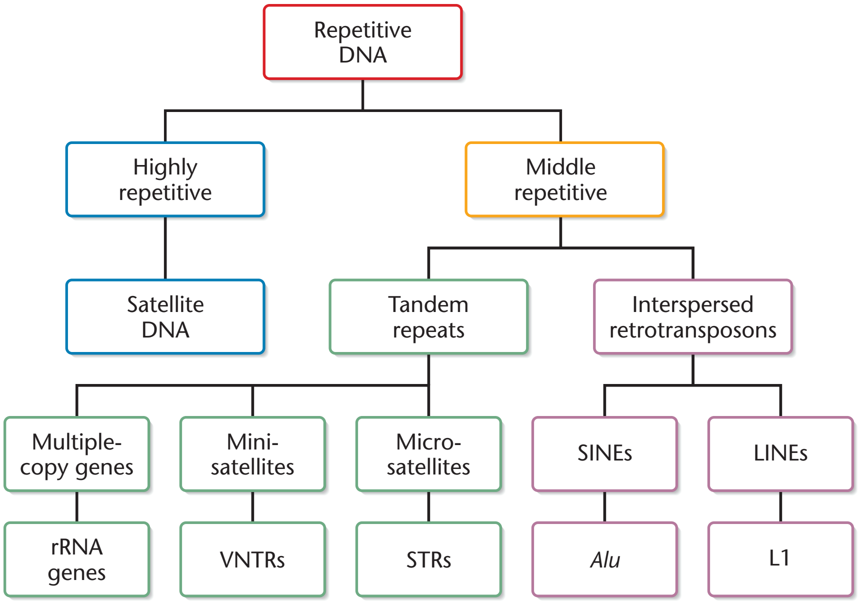


FIGURE 12.14 An overview of the various categories of repetitive DNA.

Satellite DNA

The nucleotide composition (e.g., the percentage of G-C versus A-T pairs) of the DNA of a particular species is reflected in the DNA's density, which can be measured with a technique called sedimentation equilibrium centrifugation, which in essence determines the molecule's density. When eukaryotic DNA is analyzed in this way, the majority is present and represented as a single main band, of fairly uniform density. However, one or more additional peaks indicate the presence of DNA that differs slightly in density. This component, called **satellite DNA,** makes up a variable proportion of the total DNA, depending on the species. For example, a profile of main-band and satellite DNA from the mouse is shown in **Figure 12.15**. By contrast, bacteria do not contain satellite DNA.

The significance of satellite DNA remained an enigma until the mid-1960s, when Roy Britten and David Kohne developed a technique for measuring the reassociation kinetics of DNA that had previously been dissociated into single strands. The researchers demonstrated that certain portions of DNA reannealed more rapidly than others, and concluded that rapid reannealing was characteristic of multiple DNA fragments composed of identical or nearly identical nucleotide sequences—the basis for the descriptive term *repetitive DNA*. Recall that, in contrast, bacterial DNA is nearly devoid of anything other than unique, single-copy sequences.

When satellite DNA is subjected to analysis by reassociation kinetics, it falls into the category of *highly repetitive DNA,* consisting of short sequences repeated a large number of times. Further evidence suggested that these sequences are present as tandem (meaning adjacent) repeats clustered in very specific chromosomal areas known to be heterochromatic—the regions flanking centromeres. This was discovered in 1969 when several researchers, including Mary-Lou Pardue and Joe Gall, applied the technique of ***in situ* hybridization** to the study of satellite DNA. This technique involves molecular hybridization between an isolated fraction of labeled DNA or RNA probes and the DNA contained in the chromosomes of a cytological preparation. While fluorescent probes are standard today, radioactive probes were used by Pardue and Gall. Following the hybridization procedure, autoradiography was performed to locate the chromosome areas complementary to the fraction of DNA or RNA.

Pardue and Gall demonstrated that radioactive molecular probes made from mouse satellite DNA hybridize with DNA of centromeric regions of mouse mitotic chromosomes (**Figure 12.16**). Several conclusions were drawn: Satellite DNA differs from main-band DNA in its molecular composition, as established by buoyant density studies. It is composed of short repetitive sequences. Finally, satellite DNA is found in the heterochromatic centromeric regions of chromosomes.

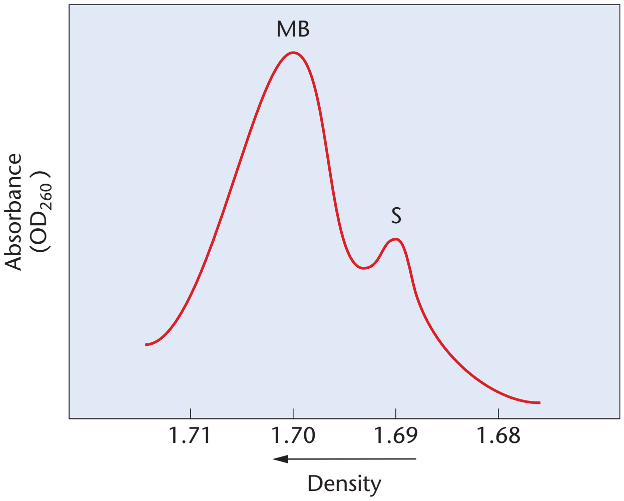


FIGURE 12.15 Separation of main-band (MB) and satellite (S) DNA from the mouse by using ultracentrifugation in a CsCl gradient.

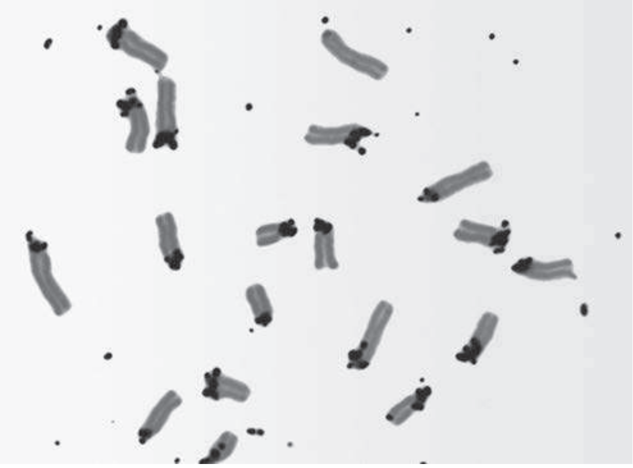


FIGURE 12.16 *In situ* hybridization between a radioactive probe representing mouse satellite DNA and mouse mitotic chromosomes. The grains in the autoradiograph are concentrated in chromosome regions referred to as the centromeres, revealing them to consist of and represent the location of satellite DNA sequences.

Centromeric DNA Sequences

The separation of homologs during mitosis and meiosis depends on **centromeres,** described cytologically in the late nineteenth century as the *primary constrictions* along eukaryotic chromosomes. In this role, it is believed that the repetitive DNA sequences contained within the centromere are critical to this role. Careful analysis has confirmed this belief. The minimal region of the centromere that supports the function of chromosomal segregation is designated the **CEN region.** Within this heterochromatic region of the chromosome, the DNA binds a platform of proteins, which in multicellular organisms includes the **kinetochore** that binds to the microtubules making up the spindle fiber during division (see Figure 2.8).

The CEN regions of the yeast *Saccharomyces cerevisiae* were the first to be studied. Each centromere serves an identical function, so it is not surprising that CENs from different chromosomes were found to be remarkably similar in their DNA sequence, displaying only minor differences between chromosomes. The CEN region of yeast chromosomes consists of about 120 bp. Mutational analysis suggests that portions near the 3′ end of this DNA region are most critical to centromere function since mutations in them, but not those nearer the 5′ end, disrupt centromere function. Thus, the DNA of this region appears to be essential to the eventual binding to the spindle fiber.

Centromere sequences of multicellular eukaryotes are much more extensive than in yeast and vary considerably in size. For example, in *Drosophila* the CEN region is found embedded within some 200–600 kb of DNA, much of which is highly repetitive (recall from our prior discussion that highly repetitive satellite DNA is localized in the centromere regions of mice). In humans, one of the most recognized satellite DNA sequences is the **alphoid family,** found mainly in the centromere regions. Alphoid sequences, each about 170 bp in length, are present in tandem arrays of up to 1 million base pairs. It is now believed that such repetitive DNA in eukaryotes is transcribed and that the RNA that is produced is ultimately involved in kinetochore function.

One final observation of interest is that the H3 histone, a normal part of most all eukaryotic nucleosomes, is substituted by a variant histone designated CENP-A in centromeric heterochromatin. It is believed that the unique N-terminal protein tails that make CENP-A unique are involved in the binding of kinetechore proteins that are essential to the microtubules of spindle fibers. This finding supports the supposition that the DNA sequence found only in centromeres is related to the function of this unique chromosomal structure.

Middle Repetitive Sequences: VNTRs and STRs

A brief look at still another prominent category of repetitive DNA sheds additional light on the organization of the eukaryotic genome. In addition to highly repetitive DNA, which constitutes about 5 percent of the human genome (and 10 percent of the mouse genome), a second category, **middle** (or **moderately**) **repetitive DNA,** recognized by reassociation kinetic studies, is fairly well characterized. Because we now know a great deal about the human genome, we will use our own species to illustrate this category of DNA in genome organization.

Although middle repetitive DNA does include some duplicated genes (such as those encoding ribosomal RNA), most prominent in this category are either noncoding tandemly repeated sequences or noncoding interspersed sequences. No function has been ascribed to these components of the genome. An example is DNA described as **variable number tandem repeats (VNTRs).** These repeating DNA sequences may be 15–100 bp long and are found within and between genes. Many such clusters are dispersed throughout the genome, and they are often referred to as **minisatellites.**

The number of tandem copies of each specific sequence at each location varies from one individual to the next, creating localized regions of 1000–20,000 bp (1–20 kb) in length. As we will see later in the text (see Chapter 21), the variation in size (length) of these regions between individual humans was originally the basis for the forensic technique referred to as **DNA fingerprinting.**

Another group of tandemly repeated sequences consists of di-, tri-, tetra-, and pentanucleotides, also referred to as **microsatellites** or **short tandem repeats (STRs).** Like VNTRs, they are dispersed throughout the genome and vary among individuals in the number of repeats present at any site. For example, in humans, the most common microsatellite is the dinucleotide $(CA)_n$, where n equals the number of repeats. Most commonly, n is between 5 and 50. These clusters have served as useful molecular markers for genome analysis.

Repetitive Transposed Sequences: SINEs and LINEs

Still another category of repetitive DNA consists of sequences that are interspersed individually throughout the genome, rather than being tandemly repeated. They can be either short or long, and many have the added distinction of being **transposable sequences,** which are mobile and can potentially move to different locations within the genome. A large portion of the human genome is composed of such sequences. Transposable sequences are discussed in more detail later in the text (see Chapter 15).

For example, **short interspersed elements,** called **SINEs,** are less than 500 base pairs long and may be present 1,500,000 times or more in the human genome. The best characterized human SINE is a set of closely related sequences called the ***Alu* family** (the name is based on the presence of DNA sequences recognized by the restriction endonuclease *Alu* I). Members of this DNA family, also found in other mammals, are 200–300 base pairs long and are dispersed rather uniformly throughout the genome, both between and within genes. In humans, the *Alu* family encompasses more than 5 percent of the entire genome.

Alu sequences are of particular interest because some members of the *Alu* family are transcribed into RNA, although the specific role of this RNA is not certain. Even so, the consequence of *Alu* sequences is their potential for transposition within the genome, which is related to chromosome rearrangements during evolution. *Alu* sequences are thought to have arisen from an RNA element whose DNA complement was dispersed throughout the genome as a result of the activity of reverse transcriptase (an enzyme that synthesizes DNA on an RNA template).

The group of **long interspersed elements (LINEs)** represents yet another category of repetitive transposable DNA sequences. LINEs are usually about 6 kb in length and in the human genome are present approximately 850,000 times. The most prominent example in humans is the **L1 family.** Members of this sequence family are about 6400 base pairs long and are present more than 500,000 times. Their 5′ end is highly variable, and their role within the genome has yet to be defined.

The general mechanism for transposition of L1 elements is now clear. The L1 DNA sequence is first transcribed into an RNA molecule. The RNA then serves as the template for synthesis of the DNA complement using the enzyme reverse transcriptase. This enzyme is encoded by a portion of the L1 sequence. The new L1 copy then integrates into the DNA of the chromosome at a new site. Because of the similarity of this transposition mechanism to that used by retroviruses, LINEs are referred to as **retrotransposons.**

SINEs and LINEs represent a significant portion of human DNA. SINEs constitute about 13 percent of the human genome, whereas LINEs constitute up to 21 percent. Within both types of elements, repeating sequences of DNA are present in combination with unique sequences.

Middle Repetitive Multiple-Copy Genes

In some cases, middle repetitive DNA includes functional genes present tandemly in multiple copies. For example, many copies exist of the genes encoding ribosomal RNA. *Drosophila* has 120 copies per haploid genome. Single genetic units encode a large precursor molecule that is processed into the 5.8*S*, 18*S*, and 28*S* rRNA components. In humans, multiple copies of this gene are clustered on the p arm of the acrocentric chromosomes 13, 14, 15, 21, and 22. Multiple copies of the genes encoding 5*S* rRNA are transcribed separately from multiple clusters found together on the terminal portion of the p arm of chromosome 1.

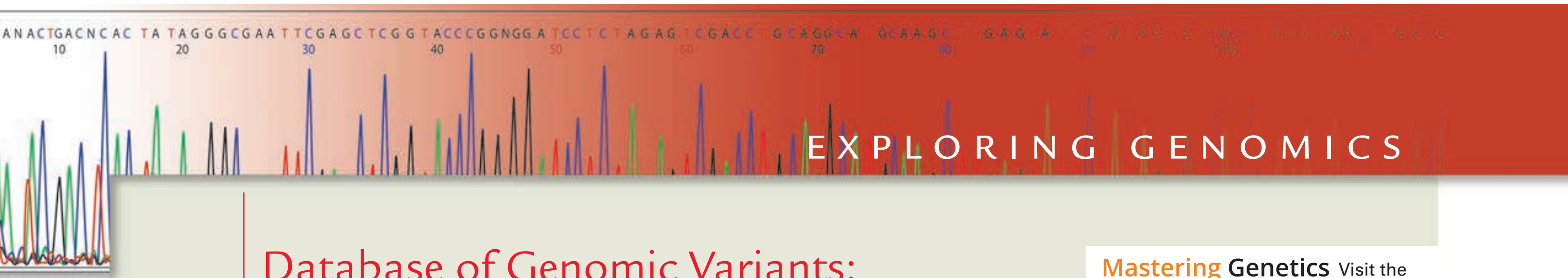
12.7 The Vast Majority of a Eukaryotic Genome Does Not Encode Functional Genes

Given the preceding information concerning various forms of repetitive DNA in eukaryotes, it is of interest to pose an important question: *What proportion of the eukaryotic genome actually encodes functional genes*?

We have seen that, taken together, the various forms of highly repetitive and moderately repetitive DNA comprise a substantial portion of the human genome—approximately 50 percent of all DNA sequences by most estimates. In addition to repetitive DNA, a large amount of the DNA consists of single-copy sequences that appear to be noncoding. Included are many instances of what we call **pseudogenes.** These are DNA sequences representing evolutionary vestiges of duplicated copies of genes that have undergone significant mutational alteration. As a result, although they show some homology to their parent gene, they are usually not transcribed because of insertions and deletions throughout their structure.

While the proportion of the genome consisting of repetitive DNA varies among eukaryotic organisms, one feature seems to be shared: *Only a very small part of the genome actually codes for proteins.* For example, the 20,000–30,000 genes encoding proteins in sea urchin occupy less than 10 percent of the genome. In *Drosophila*, only 5 to 10 percent of the genome is occupied by genes coding for proteins. In humans, it appears that the coding regions of the estimated 20,000 functional genes occupy only about 2 percent of the total DNA sequence making up the genome.

Study of the various forms of repetitive DNA has significantly enhanced our understanding of genome organization, which we will explore in more depth later in the text (see Chapter 21).



EXPLORING GENOMICS

Database of Genomic Variants: Structural Variations in the Human Genome

Mastering Genetics Visit the Study Area: Exploring Genomics

In this chapter, we focused on structural details of chromosomes and DNA sequence organization in chromosomes. A related finding is that large segments of DNA and a number of genes can vary greatly in copy number due to duplications, creating **copy number variants (CNVs).** Many studies are underway to identify and map CNVs and to find possible disease conditions associated with them.

Several thousand CNVs have been identified in the human genome, and estimates suggest there may be thousands more within human populations. In this Exploring Genomics exercise we will visit the **Database of Genomic Variants (DGV),** which provides a quickly expanding summary of structural variations in the human genome including CNVs.

Exercise I - Database of Genomic Variants

1. Access the DGV at http://dgv.tcag.ca. Click the "About the Project" tab to learn more about the purpose of the DGV.

2. Information in the DGV is easily viewed by clicking on a chromosome of interest using the "Find DGV Variants by Chromosome" feature. Using this feature, click on a chromosome of interest to you. A table will appear under the "Variants" tab showing several columns of data including:

   - Start and Stop: Shows the locus for the CNV, including the base pairs that span the variation.
   - Variant Accession: Provides a unique identifying number for each variation. Click on the variant accession number to reveal a separate page of specific details about each CNV, including the chromosomal banding location for the variation and known genes that are located in the CNV.
   - Variant Type: Most variations in this database are CNVs. Variant subtypes, such as deletions or insertions, and duplications, are shown in an adjacent column.

3. Let's analyze a particular group of CNVs. Many CNVs are unlikely to affect phenotype because they involve large areas of non-protein-coding or nonregulatory sequences. But gene-containing CNVs have been identified, including variants containing genes associated with Alzheimer disease, Parkinson disease, and other conditions.

   Defensin *(DEF)* genes are part of a large family of highly duplicated genes. To learn more about *DEF* genes and CNVs, use the Keyword Search box and search for *DEF.* A results page for the search will appear with a listing of relevant CNVs. Click on the name for any of the different *DEF* genes listed, which will take you to a wealth of information [including links to Online Mendelian Inheritance in Man (OMIM)] about these genes so that you can answer the following questions. Do this for several *DEF*-containing CNVs on different chromosomes to find the information you will need for your answers.

   a. On what chromosome(s) did you find CNVs containing *DEF* genes?

   b. What did you learn about the function of *DEF* gene products? What do DEF proteins do?

   c. Variations in *DEF* genotypes and *DEF* gene expression in humans have been implicated in a number of different human disease conditions. Give examples of the kinds of disorders affected by variations in *DEF* genotypes.

   d. Explore the DGV to search a chromosome of interest to you and learn more about CNVs that have been mapped to that chromosome. Try the Genome Browser feature that will show you maps of each chromosome indicating different variations. For CNVs (shown in blue), clicking on the CNV will take you to its locus on the chromosome.

CASE STUDY Helping or hurting?

Roberts syndrome is a rare inherited disorder characterized by facial defects as well as severe limb shortening, extra digits, and deformities of the knees and ankles. A cytogenetic analysis of patients with Roberts syndrome, using Giemsa staining or C-banding, reveals that there is premature separation of centromeres and other heterochromatic regions during mitotic metaphase instead of anaphase. A couple with an affected infant is contacted by a local organization dedicated to promoting research on rare genetic diseases, asking if they can photograph the infant as part of a campaign to obtain funding for these conditions. The couple learned that the privacy of such medical images is not well protected, and they often are subsequently displayed on public websites. The couple was torn between helping to raise awareness and promoting research on this condition and sheltering their child from having his images used inappropriately. Several interesting questions are raised.

1. In Roberts syndrome, how could premature separation of centromeres during mitosis cause the wide range of phenotypic deficiencies?
2. What ethical obligations do the parents owe to their child in this situation and to helping others with Roberts syndrome by allowing images of their child to be used in raising awareness of this disorder?
3. If the parents decide to allow their infant to be photographed, what steps should the local organization take to ensure appropriate use and distribution of the photos?

See Onion R. (2014). History, or Just Horror? *Slate Magazine* (http://www.slate.com/articles/news_and_politics/history/2014/11/old_medical_photographs_are_images_of_syphilis_and_tuberculosis_patients.html).

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Bacteriophage and bacterial chromosomes, in contrast to those of eukaryotes, are largely devoid of associated proteins, are of much smaller size, and, with some exceptions, consist of circular DNA. Supercoiling facilitates the compaction of the DNA of these chromosomes.
2. Polytene and lampbrush chromosomes are examples of specialized structures that extended our knowledge of genetic organization and function well in advance of the technology available to the modern-day molecular biologist.
3. Eukaryotic chromatin is a nucleoprotein organized into repeating units called nucleosomes, which are composed of about 200 base pairs of DNA, an octamer of four types of histones, plus one linker histone.
4. Nucleosomes provide a mechanism for compaction of chromatin within the nucleus. Several forms of chemical modification, for example, acetylation and methylation, may alter the level of compaction, a process referred to as chromatin remodeling, which is critical to replication and transcription of DNA.
5. Heterochromatin, prematurely condensed in interphase and for the most part genetically inert, is illustrated by centromeric and telomeric regions of eukaryotic chromosomes, the Y chromosome, and the Barr body.
6. Chromosome banding techniques provide a way to subdivide and identify specific regions of mitotic chromosomes.
7. Eukaryotic genomes demonstrate complex sequence organization characterized by numerous categories of repetitive DNA, consisting of either tandem repeats clustered in various regions of the genome or single sequences repeatedly interspersed at random throughout the genome.
8. The vast majority of the DNA in most eukaryotic genomes does not encode functional genes. In humans, for example, only about 2 percent of the genome is used to encode the 20,000 genes found there.

INSIGHTS AND SOLUTIONS

A previously undiscovered single-celled organism was found living at a great depth on the ocean floor. Its nucleus contains only a single, linear chromosome consisting of 7×10^6 nucleotide pairs of DNA coalesced with three types of histone-like proteins. Consider the following questions:

1. A short micrococcal nuclease digestion yielded DNA fractions consisting of 700, 1400, and 2100 base pairs. Predict what these fractions represent. What conclusions can be drawn?

Solution: The chromatin fiber may consist of a nucleosome variant containing 700 base pairs of DNA. The 1400- and 2100-bp fractions represent two and three of these nucleosomes, respectively, linked together. Enzymatic digestion

(*continued*)

Insights and Solutions—continued

may have been incomplete, leading to the latter two fractions.

2. The analysis of individual nucleosomes revealed that each unit contained one copy of each protein and that the short linker DNA had no protein bound to it. If the entire chromosome consists of nucleosomes (discounting any linker DNA), how many are there, and how many total proteins are needed to form them?

   Solution: Since the chromosome contains 7×10^6 base pairs of DNA, the number of nucleosomes, each containing 7×10^2 base pairs, is equal to

   $$7 \times 10^6/7 \times 10^2 = 10^4 \text{ nucleosomes}$$

   The chromosome thus contains 10^4 copies of each of the three proteins, for a total of 3×10^4 proteins.

3. Analysis then revealed the organism's DNA to be a double helix similar to the Watson–Crick model, but containing 20 base pairs per complete turn of the right-handed helix. The physical size of the nucleosome was exactly double the volume occupied by the nucleosome found in any other known eukaryote, and the nucleosome's axis length was greater by a factor of two. Compare the degree of compaction (the number of turns per nucleosome) of this organism's nucleosome with that found in other eukaryotes.

   Solution: The unique organism compacts a length of DNA consisting of 35 complete turns of the helix (700 base pairs per nucleosome/20 base pairs per turn) into each nucleosome. The normal eukaryote compacts a length of DNA consisting of 20 complete turns of the helix (200 base pairs per nucleosome/10 base pairs per turn) into a nucleosome half the volume of that in the unique organism. The degree of compaction is therefore less in the unique organism.

4. No further coiling or compaction of this unique chromosome occurs in the newly discovered organism. Compare this situation with that of a eukaryotic chromosome. Do you think an interphase human chromosome 7×10^6 base pairs in length would be a shorter or longer chromatin fiber?

   Solution: The eukaryotic chromosome contains still another level of condensation in the form of solenoids, which are coils consisting of nucleosomes connected with linker DNA. Solenoids condense the eukaryotic fiber by still another factor of five. The length of the unique chromosome is compacted into 10^4 nucleosomes, each containing an axis length twice that of the eukaryotic fiber. The eukaryotic fiber consists of $7 \times 10^6/2 \times 10^2 = 3.5 \times 10^4$ nucleosomes, 3.5 times more than the unique organism. However, they are compacted by the factor of five in each solenoid. Therefore, the chromosome of the unique organism is a longer chromatin fiber.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. HOW DO WE KNOW? In this chapter, we focused on how DNA is organized at the chromosomal level. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
(a) How do we know that viral and bacterial chromosomes most often consist of circular DNA molecules devoid of protein?
(b) What is the experimental basis for concluding that puffs in polytene chromosomes and loops in lampbrush chromosomes are areas of intense transcription of RNA?
(c) How did we learn that eukaryotic chromatin exists in the form of repeating nucleosomes, each consisting of about 200 base pairs and an octamer of histones?
(d) How do we know that satellite DNA consists of repetitive sequences and has been derived from regions of the centromere?

2. CONCEPT QUESTION Review the Chapter Concepts list on p. 263. These all relate to how DNA is organized in viral, bacterial, and eukaryote chromosomes. Write a short essay that contrasts the major differences between the organization of DNA in viruses and bacteria versus eukaryotes.

3. Contrast the size of the single chromosome in bacteriophage λ and T2 with that of *E. coli*. How does this relate to the relative size and complexity of phages and bacteria?

4. Describe the structure of giant polytene chromosomes and how they arise.

5. What genetic process is occurring in a puff of a polytene chromosome? How do we know this experimentally?

6. During what genetic process are lampbrush chromosomes present in vertebrates?

7. Why might we predict that the organization of eukaryotic genetic material will be more complex than that of viruses or bacteria?

8. Describe the sequence of research findings that led to the development of the model of chromatin structure.

9. Describe the molecular composition and arrangement of the components in the nucleosome.

10. Describe the transitions that occur as nucleosomes are coiled and folded, ultimately forming a chromatid.

11. Provide a comprehensive definition of heterochromatin and list as many examples as you can.

12. Mammals contain a diploid genome consisting of at least 10^9 bp. If this amount of DNA is present as chromatin fibers, where each group of 200 bp of DNA is combined with 9 histones into a nucleosome and each group of 6 nucleosomes is combined into a solenoid, achieving a final packing ratio of 50, determine (a) the total number of nucleosomes in all fibers, (b) the total number of histone molecules combined with DNA in the diploid genome, and (c) the combined length of all fibers.

13. Assume that a viral DNA molecule is a 50-μm-long circular strand with a uniform 20-Å diameter. If this molecule is contained in a viral head that is a 0.08-μm-diameter sphere, will the DNA molecule fit into the viral head, assuming complete flexibility of the molecule? Justify your answer mathematically.

14. How many base pairs are in a molecule of phage T2 DNA 52 μm long?

15. Examples of histone modifications are acetylation (by histone acetyltransferase, or HAT), which is often linked to gene activation, and deacetylation (by histone deacetylases, or HDACs), which often leads to gene silencing typical of heterochromatin. Such heterochromatinization is initiated from a nucleation site and spreads bidirectionally until encountering boundaries that delimit the silenced areas. Recall from earlier in the text (see Chapter 4) the brief discussion of position effect, where repositioning of the w^+ allele in *Drosophila* by translocation or inversion near heterochromatin produces intermittent w^+ activity. In the heterozygous state (w^+/w), a variegated eye is produced, with white and red patches. How might one explain position-effect variegation in terms of histone acetylation and/or deacetylation?

16. Contrast the structure of SINE and LINE DNA sequences. Why are LINEs referred to as retrotransposons?

17. Variable number tandem repeats (VNTRs) are repeating DNA sequences of about 15–100 bp in length, found both within and between genes. Why are they commonly used in forensics?

18. It has been shown that infectious agents such as viruses often exert a dramatic effect on their host cell's genome architecture. In many cases, viruses induce methylation of host DNA sequences in order to enhance their infectivity. What specific host gene functions would you consider as strong candidates for such methylation by infecting viruses?

19. Cancer can be defined as an abnormal proliferation of cells that defy the normal regulatory controls observed by normal cells. Recently, histone deacetylation therapies have been attempted in the treatment of certain cancers [reviewed by Delcuve et al. (2009)]. Specifically, the FDA has approved histone deacetylation (HDAC) inhibitors for the treatment of cutaneous T-cell lymphoma. Explain why histone acetylation might be associated with cancer and what the rationale is for the use of HDAC inhibitors in the treatment of certain forms of cancer.

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

20. In a study of *Drosophila,* two normally active genes, w^+ (wild-type allele of the *white*-eye gene) and *hsp*26 (a heat-shock gene), were introduced (using a plasmid vector) into euchromatic and heterochromatic chromosomal regions, and the relative activity of each gene was assessed [Sun et al. (2002)]. An approximation of the resulting data is shown in the following table. Which characteristic or characteristics of heterochromatin are supported by the experimental data?

	Activity (relative percentage)	
Gene	Euchromatin	Heterochromatin
***hsp*26**	100%	31%
w^+	100%	8%

21. While much remains to be learned about the role of nucleosomes and chromatin structure and function, recent research indicates that *in vivo* chemical modification of histones is associated with changes in gene activity. One study determined that acetylation of H3 and H4 is associated with 21.1 percent and 13.8 percent increases in yeast gene activity, respectively, and that histones associated with yeast heterochromatin are hypomethylated relative to the genome average [Bernstein et al. (2000)]. Speculate on the significance of these findings in terms of nucleosome–DNA interactions and gene activity.

22. An article entitled "Nucleosome Positioning at the Replication Fork" states: "both the 'old' randomly segregated nucleosomes as well as the 'new' assembled histone octamers rapidly position themselves (within seconds) on the newly replicated DNA strands" [Lucchini et al. (2002)]. Given this statement, how would one compare the distribution of nucleosomes and DNA in newly replicated chromatin? How could one experimentally test the distribution of nucleosomes on newly replicated chromosomes?

23. The human genome contains approximately 10^6 copies of an *Alu* sequence, one of the best-studied classes of short interspersed elements (SINEs), per haploid genome. Individual *Alu* units share a 282-nucleotide consensus sequence followed by a 3′-adenine-rich tail region [Schmid (1998)]. Given that there are approximately 3×10^9 base pairs per human haploid genome, about how many base pairs are spaced between each *Alu* sequence?

24. Following is a diagram of the general structure of the bacteriophage λ chromosome. Speculate on the mechanism by which it forms a closed ring upon infection of the host cell.

```
5′ GGGCGGCGACCT—double-stranded region—3′
                        3′—double-stranded region—CCCGCCGCTGGA 5′
```

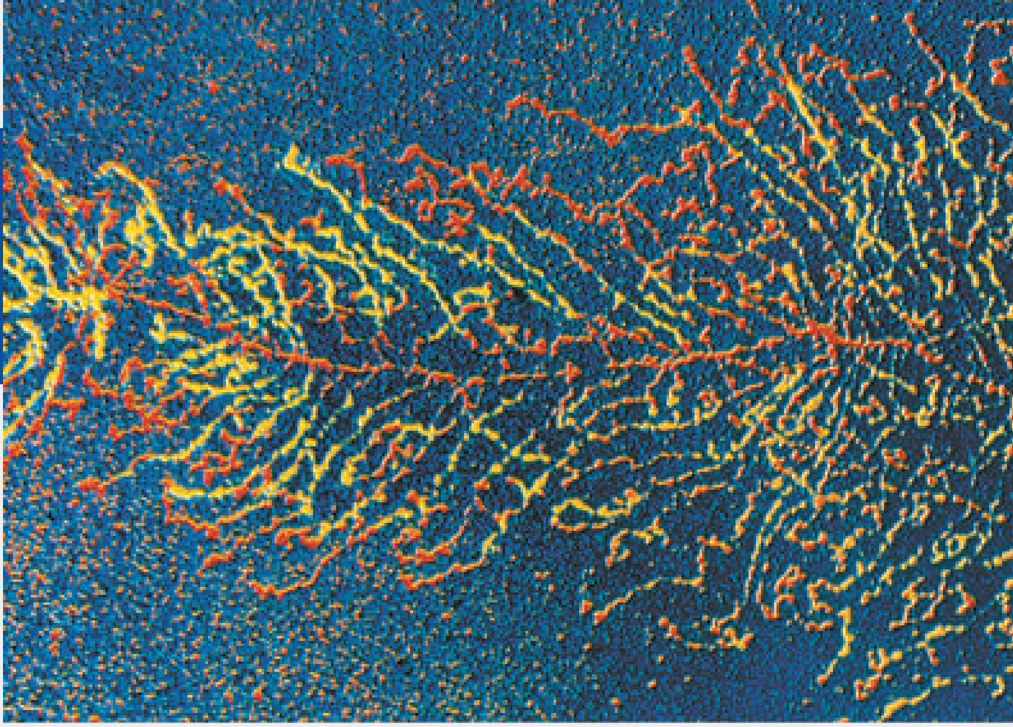
25. Microsatellites are currently exploited as markers for paternity testing. A sample paternity test is shown in the following table in which ten microsatellite markers were used to test samples from a mother, her child, and an alleged father. The name of the microsatellite locus is given in the left-hand column, and the genotype of each individual is recorded as the number of repeats he or she carries at that locus. For example, at locus D9S302, the mother carries 30 repeats on one of her chromosomes and 31 on the other. In cases where an individual carries the same number of repeats on both chromosomes, only a single number is recorded. (Some of the numbers are followed by a decimal point, for example, 20.2, to indicate a partial repeat in addition to the complete repeats.) Assuming that these markers are inherited in a simple Mendelian fashion, can the alleged father be excluded as the source of the sperm that produced the child? Why or why not? Explain.

Microsatellite Locus-Chromosome Location	Mother	Child	Alleged Father
D9S302-9q31-q33	30	31	32
	31	32	33
D22S883-22pter-22qter	17	20.2	20.2
	22	22	
D18S535-18q12.2-q12.3	12	13	11
	14	14	13
D7SI804-7pter-7qter	27	26	26
	30	30	27
D3S2387-3p24.2.3pter	23	24	20.2
	25.2	25.2	24
D4S2386-4pter-qter	12	12	12
			16
D5S1719-5pter-5qter	11	10.3	10
	11.3	11	10.3
CSF1PO-5q33.3.q34	11	11	10
		12	12
FESFPS-15q25-15qter	11	12	10
	12	13	13
TH01-11p15.5	7	7	7
			8

26. At the end of the short arm of human chromosome 16 (16p), several genes associated with disease are present, including thalassemia and polycystic kidney disease. When that region of chromosome 16 was sequenced, gene-coding regions were found to be very close to the telomere-associated sequences. Could there be a possible link between the location of these genes and the presence of the telomere-associated sequences? What further information concerning the disease genes would be useful in your analysis?

27. Spermatogenesis in mammals results in sperm that have a nucleus that is 40 times smaller than an average somatic cell. Thus, the sperm haploid genome must be packaged very tightly, yet in a way that is reversible after fertilization. This sperm-specific DNA compaction is due to a nucleosome-to-nucleoprotamine transition, where the histone-based nucleosomes are removed and replaced with arginine-rich protamine proteins that facilitate a tighter packaging of DNA. In 2013 Montellier et al. showed that replacement of the H2B protein in the nucleosomes with a testis-specific variant of H2B called TSH2B is a critical step prior to the nucleosome-to-nucleoprotamine transition. Mice lacking TSH2B retain H2B and their sperm arrest late in spermatogenesis with reduced DNA compaction. Based on these findings, would you expect that TSH2B-containing nucleosomes are more or less stable than H2B-containing nucleosomes? Explain your reasoning.

13



The Genetic Code and Transcription

Electron micrograph of a segment of DNA undergoing transcription.

CHAPTER CONCEPTS

- Genetic information is stored in DNA and encoded in a form that is nearly universal in all living things on Earth.
- The genetic code is initially transferred from DNA to RNA, in the process of transcription. Once transferred to RNA, the genetic code exists as triplet codons, which are sets of three ribonucleotides in which each ribonucleotide is one of the four kinds composing RNA.
- RNA's four ribonucleotides, analogous to an alphabet of four "letters," can be arranged into 64 different three-letter sequences. Most of these triplet codons in RNA encode one of the 20 amino acids present in proteins.
- Several codons act as signals that initiate or terminate protein synthesis.
- In bacteria, the process of transcription is similar to, but less complex, than in eukaryotes, where the initial transcript must be processed prior to its translation.

As we saw earlier in the text (see Chapter 10), the structure of DNA consists of a linear sequence of deoxyribonucleotides. This sequence ultimately dictates the composition of proteins, the end products of protein-coding genes. A central issue is how information stored as a nucleic acid can be decoded into a protein. Figure 13.1 provides a simple overview of how this transfer of information, resulting ultimately in gene expression, occurs. In the first step, information present on one of the two strands of DNA (the template strand) is transferred into an RNA complement through the process of transcription. Once synthesized, this RNA acts as a "messenger" molecule, which in eukaryotes is transported out of the nucleus. The mRNAs then associate with ribosomes, where decoding into proteins occurs. The directional flow of genetic information from DNA to RNA to protein is called the **central dogma of molecular genetics**.

In this chapter, we will focus on the initial phases of gene expression by addressing two major questions. First, how is genetic information encoded? Second, how does the transfer of information from DNA to RNA occur, thus explaining the process of transcription? As you will see, ingenious analytical research has established that the genetic code is written in units of three letters—triplets of ribonucleotides in mRNA that reflect the stored information in genes. Most of the triplets direct the incorporation of a specific amino acid into a protein as it is synthesized. As we can predict, based on the complexity of the replication of DNA (see Chapter 11), transcription is also a complex process, dependent on a major polymerase enzyme and a cast of supporting proteins. We will explore what is known about transcription in

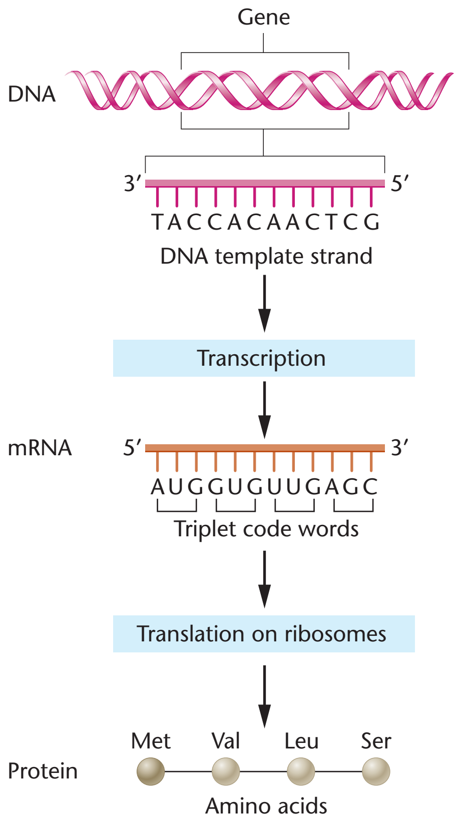


FIGURE 13.1 The central dogma. Gene expression consists of transcription of DNA into mRNA (top) and the translation (center) of mRNA (with the help of a ribosome) into protein (bottom).

bacteria and then contrast this bacterial model with transcription in eukaryotes. In Chapter 14 we will continue our discussion of gene expression by addressing how translation occurs and then describing the structure and function of proteins. Together, these chapters provide a comprehensive picture of molecular genetics, which serves as the most basic foundation for understanding living organisms.

13.1 The Genetic Code Uses Ribonucleotide Bases as "Letters"

Before we consider the various analytical approaches that led to our current understanding of the genetic code, let's summarize the general features that characterize it:

1. The genetic code is written in linear form, using as "letters" the bases in ribonucleotides that compose mRNA molecules. The ribonucleotide sequence is derived from the complementary nucleotide bases in DNA.
2. Each "word" within an mRNA consists of three ribonucleotide letters, thus referred to as a **triplet code**.
3. With several exceptions, each group of three ribonucleotides, called a **codon,** specifies one amino acid, making the code nearly **unambiguous**.
4. The code is **degenerate,** meaning that a given amino acid can be specified by more than one triplet codon. This is the case for 18 of the 20 amino acids.
5. The code contains one "start" and three "stop" signals, triplets that **initiate** and **terminate** translation, respectively.
6. No internal punctuation (analogous, for example, to a comma) is used in the code. Thus, the code is said to be **commaless.** Once translation of mRNA begins, the codons are read one after the other with no breaks between them (until a stop signal is reached).
7. The code is **nonoverlapping.** After translation commences, any single ribonucleotide within the mRNA is part of only one triplet.
8. The code is **colinear;** the order of codons in the mRNA determines the order of amino acids in the encoded protein.
9. The code is nearly **universal.** With only minor exceptions, a single coding dictionary is used by almost all viruses, bacteria, archaea, and eukaryotes.

13.2 Early Studies Established the Basic Operational Patterns of the Code

In the late 1950s, before it became clear that mRNA is the intermediate that transfers genetic information from DNA to proteins, researchers thought that DNA itself might directly encode proteins during their synthesis. Because ribosomes had already been identified, the initial thinking was that information in DNA was transferred to the RNA of the ribosome, which served as the template for protein synthesis in the cytoplasm. This concept was soon recognized as untenable as accumulating evidence indicated the existence of an unstable intermediate template. The RNA of ribosomes, on the other hand, was extremely stable. As a result, in 1961 François Jacob and Jacques Monod postulated the existence of **messenger RNA (mRNA).** Once mRNA was discovered, it was clear that even though genetic information is stored in DNA, the code that is translated into proteins actually resides in RNA. The central question then was how only four letters—the four ribonucleotides—could spell enough words to specify the 20 amino acids.

The Triplet Nature of the Code

In the early 1960s, Sydney Brenner argued on theoretical grounds that the code had to be triplet based since three-letter words represent the minimal use of four letters to specify 20 amino acids. A code of four nucleotides, taken two at a time, for example, provides only 16 unique code words (4^2). A triplet code yields 64 words (4^3)—clearly more than the 20 needed—and is much simpler than a four-letter code (4^4), which specifies 256 words.

Experimental evidence supporting the triplet nature of the code was subsequently derived from research by Francis Crick and his colleagues. Using phage T4, they studied **frameshift mutations,** which result from the addition or deletion of one or more nucleotides within a gene and subsequently the mRNA transcribed from it. The gain or loss of letters shifts the **reading frame** during translation. Crick and his colleagues found that the gain or loss of one or two nucleotides caused a frameshift mutation, but when three nucleotides were involved, the frame of reading was reestablished (**Figure 13.2**). This would not occur if the code was anything other than a triplet.

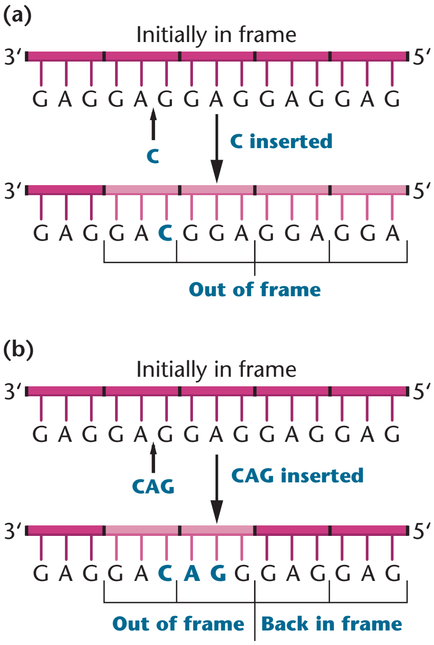


FIGURE 13.2 The effect of frameshift mutations on a DNA sequence with the repeating triplet sequence GAG. (a) The insertion of a single nucleotide shifts all subsequent triplet reading frames. (b) The insertion of three nucleotides changes only two triplets, but the frame of reading is then reestablished to the original sequence.

13.3 Studies by Nirenberg, Matthaei, and Others Led to Deciphering of the Code

In 1961, Marshall Nirenberg and J. Heinrich Matthaei became the first to link specific coding sequences to specific amino acids, laying a cornerstone for the complete analysis of the genetic code. Their success, as well as that of others who made important contributions in deciphering the code, was dependent on the use of two experimental tools: (1) a cell-free (*in vitro*) system for synthesizing proteins and (2) a means of producing synthetic mRNAs to serve as templates for polypeptide synthesis in the cell-free system.

Synthesizing Polypeptides in a Cell-Free System

In a cell-free protein-synthesizing system, amino acids are incorporated into polypeptide chains in a test tube (hence *in vitro*, literally "in glass"). The process begins with a cell extract, or lysate, containing all the essential factors for protein synthesis: ribosomes, tRNAs, amino acids, and other molecules essential to translation (see Chapter 14), but with organelles and cell membranes removed. To allow scientists to follow (or "trace") the progress of protein synthesis, one or more of the amino acids must be radioactive. Finally, an mRNA must be added to serve as the template to be translated.

In 1961, cellular mRNA had yet to be isolated. However, use of the enzyme **polynucleotide phosphorylase** allowed the artificial synthesis of RNA templates, which could be added to the cell-free system. This enzyme, isolated from bacteria, catalyzes the reaction shown in **Figure 13.3**. Discovered in 1955 by Marianne Grunberg-Manago and Severo Ochoa, the enzyme functions metabolically in bacterial cells to degrade RNA. However, *in vitro,* in the presence of high concentrations of ribonucleoside diphosphates, the reaction can be "forced" in the opposite direction, to synthesize RNA, as illustrated in Figure 13.3.

In contrast to RNA polymerase (which constructs mRNA *in vivo*), polynucleotide phosphorylase does not require a DNA template. As a result, the order in which ribonucleotides are added to a growing RNA is random, depending only on the relative concentrations of the four ribonucleoside diphosphates present in the reaction mixture. The probability of the insertion of a specific ribonucleotide is proportional to the availability of that molecule relative to other available ribonucleotides. *This point is absolutely critical to understanding the work of Nirenberg and others in the ensuing discussion.*

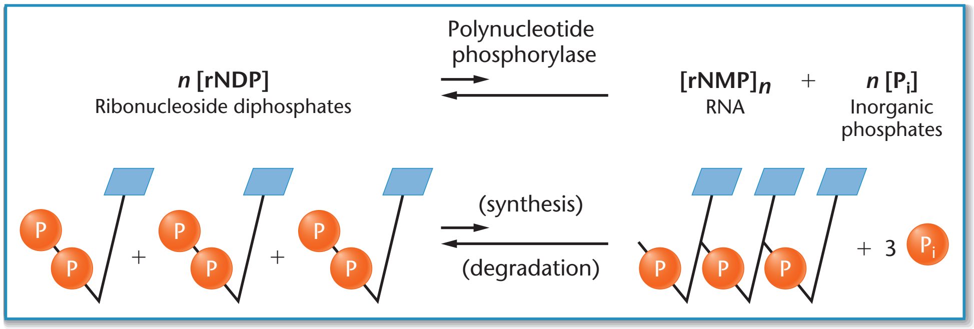


FIGURE 13.3 The reaction catalyzed by the enzyme polynucleotide phosphorylase. Note that the equilibrium of the reaction favors the degradation of RNA but that the reaction can be “forced” in the direction favoring synthesis.

Together, the cell-free system for protein synthesis and the availability of synthetic mRNAs provided a means of deciphering the ribonucleotide composition of various codons encoding specific amino acids.

Homopolymer Codes

For their initial experiments, Nirenberg and Matthaei synthesized **RNA homopolymers,** RNA molecules containing only one type of ribonucleotide. In other words, the mRNA they added to their cell-free protein-synthesizing system was either UUUUUU . . . , AAAAAA . . . , CCCCCC . . . , or GGGGGG . . . . They tested each of these types of mRNA to see which, if any, amino acids were consequently incorporated into newly synthesized proteins. Multiple experimental syntheses were conducted with each homopolymer. They always made all 20 amino acids available, but for each experiment they attached a radioactive label to a different amino acid and thus could tell when that amino acid had been incorporated into the resulting polypeptide.

For example, from experiments using ^{14}C-phenylalanine (**Table 13.1**), Nirenberg and Matthaei concluded that the RNA homopolymer UUUUU . . . (polyuridylic acid or poly U) directed the incorporation of only phenylalanine into the peptide homopolymer polyphenylalanine. Assuming the validity of a triplet code, they made the first specific codon assignment: UUU codes for phenylalanine. Running similar experiments, they quickly found that AAA codes for lysine and CCC codes for proline. Poly G was not a functional template, probably because the molecule folds back on itself, likely blocking association with the ribosome. Thus, the assignment for GGG had to await other approaches.

TABLE 13.1 Incorporation of ^{14}C-Phenylalanine into Protein

Artificial mRNA	Radioactivity (counts/min)
None	44
Poly U	39,800
Poly A	50
Poly C	38

Source: After M. Nirenberg and J. H. Matthaei (1961).

Note that the specific triplet codon assignments were possible only because homopolymers were used. In this method, only the general nucleotide composition of the template is known, not the specific order of the nucleotides in each triplet. But since three identical letters can have only one possible sequence (e.g., UUU), three of the actual codons for phenylalanine, lysine, and proline could be identified.

The Use of Mixed Heteropolymers

With the initial success of work with homopolymers, Nirenberg and Matthaei, and Ochoa and coworkers, turned to the use of mixed **RNA heteropolymers** in which two, three, or four different ribonucleoside diphosphates were used in combination to form the artificial message. The researchers reasoned that if they knew the relative proportion of each type of ribonucleoside diphosphate in their synthetic mRNA, they could predict the frequency of each of the possible triplet codons it contained. If they then added the mRNA to the cell-free system and ascertained the percentage of each amino acid present in the resulting polypeptide, they could analyze the results and predict the *composition* of the triplets that had specified those particular amino acids.

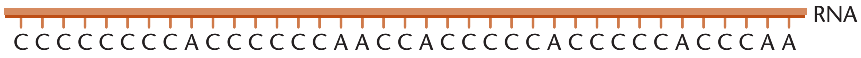
This approach is illustrated in **Figure 13.4**. Suppose that only A and C are used for synthesizing the mRNA, in a ratio of 1A : 5C. The insertion of a ribonucleotide at any position along the RNA molecule during its synthesis is determined by the ratio of A : C. Therefore, there is a 1/6 possibility for an A and a 5/6 chance for a C to occupy each position. On this basis, we can calculate the frequency of any given triplet appearing in the message.

For AAA, the frequency is $(1/6)^3$ or about 0.5 percent. For AAC, ACA, and CAA, the frequencies are identical—that is, $(1/6)^2(5/6)$ or about 2.3 percent for each triplet. Together, all three 2A : 1C triplets account for 6.9 percent of the total three-letter sequences. In the same way, each of three 1A : 2C triplets accounts for $(1/6)(5/6)^2$ or 11.6 percent (or a

RNA Heteropolymer with Ratio of 1A:5C

Possible compositions	Possible triplets	Probability of occurrence of any triplet	Final %
3A	AAA	$(1/6)^3 = 1/216 = 0.5\%$	0.5
2A:1C	AAC ACA CAA	$(1/6)^2(5/6) = 5/216 = 2.3\%$	$3 \times 2.3 = 6.9$
1A:2C	ACC CAC CCA	$(1/6)(5/6)^2 = 25/216 = 11.6\%$	$3 \times 11.6 = 34.8$
3C	CCC	$(5/6)^3 = 125/216 = 57.9\%$	57.9
			~100

Chemical synthesis of message


RNA
C C C C C C C C A C C C C C C A A C A C C C C C A C C C C C A C C C A A

Translation of message

Percentage of amino acids in protein		Probable base-composition assignments
Lysine	<1	AAA
Glutamine	2	2A:1C
Asparagine	2	2A:1C
Threonine	12	1A:2C
Histidine	14	1A:2C, 2A:1C
Proline	69	CCC, 1A:2C

FIGURE 13.4 Results and interpretation of a heteropolymer experiment in which a ratio of 1A : 5C (1/6A : 5/6C) is used.

total of 34.8 percent); CCC is represented by $(5/6)^3$, or 57.9 percent of all triplets.

By examining the percentages of the different amino acids incorporated into the polypeptide synthesized under the direction of the mRNA just described, we can propose probable base compositions for the codons specifying each of those amino acids (bottom of Figure 13.4). Because proline appears 69 percent of the time, we could propose that proline is encoded by CCC (57.9 percent) and also by one of the codons consisting of 1A : 2C (11.6 percent). Histidine, at 14 percent, is probably coded by one 1A : 2C codon (11.6 percent) and one 2A : 1C codon (2.3 percent). Threonine, at 12 percent, is likely coded by only one 1A : 2C codon. Asparagine and glutamine each appear to be coded by one of the 2A : 1C codons, and lysine appears to be coded by AAA.

Using as many as all four ribonucleotides to construct mRNAs, the researchers conducted many similar experiments. Although the determination by this means of the *composition* of triplet code words corresponding to all 20 amino acids represented a very significant breakthrough, the *specific sequences* of triplets were still unknown—other approaches were still needed.

NOW SOLVE THIS

13.1 In a mixed heteropolymer experiment using polynucleotide phosphorylase, 3/4G : 1/4C was used to form the synthetic message. The amino acid composition of the resulting protein was determined to be:

Glycine	36/64	56 percent
Alanine	12/64	19 percent
Arginine	12/64	19 percent
Proline	4/64	6 percent

From this information,

(a) Indicate the percentage (or fraction) of the time each possible codon will occur in the message.

(b) Determine one consistent codon base composition assignment for the amino acids present.

■ **HINT:** *This problem asks you to analyze a heteropolymer experiment and to predict codon composition assignments for the amino acids encoded by the synthetic message. The key to its solution is to first calculate the proportion of each triplet codon in the synthetic RNA and then match these to the proportions of amino acids that are synthesized.*

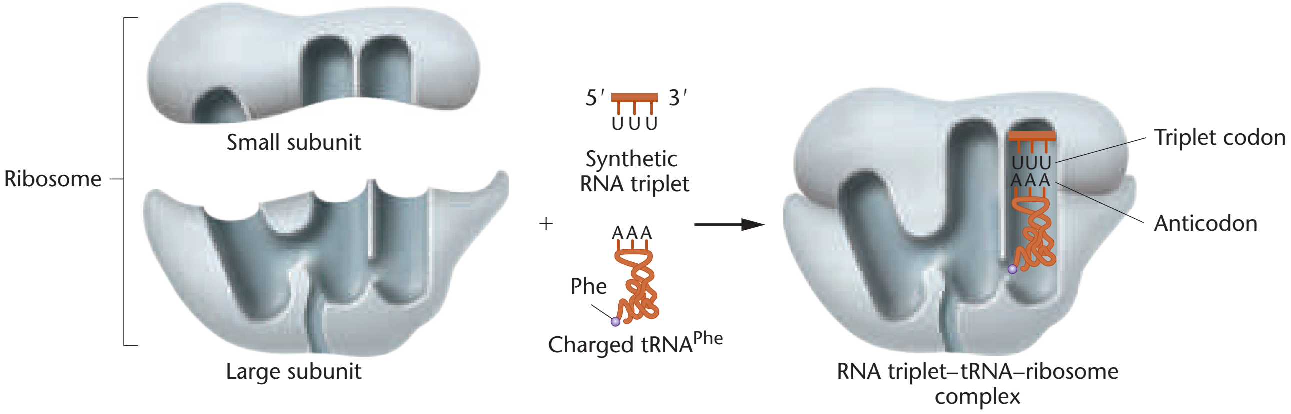


FIGURE 13.5 The behavior of the components during the triplet binding assay. When the synthetic UUU triplet is positioned in the ribosome, it acts as a codon, attracting the complementary AAA anticodon of the charged tRNA^{Phe}.

The Triplet Binding Assay

It was not long before more advanced techniques for elucidating codons were developed. In 1964, Nirenberg and Philip Leder developed the **triplet binding assay,** leading to specific sequence assignments for triplet codons. This technique took advantage of the observation that ribosomes, when presented *in vitro* with an RNA sequence as short as three ribonucleotides, will bind to it and form a complex similar to what is found *in vivo.* The triplet RNA sequence acts like a codon in mRNA, attracting a **transfer RNA (tRNA)** molecule containing a complementary sequence and carrying a specific amino acid (Figure 13.5). Such a triplet sequence in tRNA, that is, complementary to a codon of mRNA, is known as an **anticodon**.

Although it was not yet feasible to chemically synthesize long stretches of RNA in the laboratory, specific triplet sequences could be synthesized to serve as templates. All that was needed was a method to determine which tRNA–amino acid was bound to the triplet RNA–ribosome complex.

The test system Nirenberg and Leder devised was quite simple. The amino acid to be tested was made radioactive and added to the cell lysate. Enzymes in the lysate (to be discussed in Chapter 14) attached the radioactive amino acid to its cognate tRNA, creating a "charged" tRNA. Because codon compositions (though not exact sequences) were known, it was possible to narrow the decision as to which amino acids should be tested for each specific triplet.

The radioactively charged tRNA, the RNA triplet, and ribosomes were incubated together and then applied to a nitrocellulose filter. The filter retains ribosomes (because of their larger size) but not the other, smaller components, such as charged tRNA. If radioactivity was not retained on the filter, an incorrect amino acid had been tested. If radioactivity remained on the filter, it did so because the charged tRNA had bound to the RNA triplet associated with the ribosome, which itself remained on the filter. In such a case, a specific codon assignment could be made.

Work proceeded in several laboratories, and in many cases clear-cut, unambiguous results were obtained. For example, **Table 13.2** shows 26 triplet codons assigned to ten different amino acids. However, in some cases the triplet binding was inefficient, and assignments were not possible. Eventually, about 50 of the 64 possible triplets were assigned. These specific assignments led to two major conclusions. First, the genetic code is *degenerate*—that is, one amino acid may be specified by more than one triplet. Second, the code is also *unambiguous*—that is, a single codon specifies only one amino acid. As we will see later in this chapter, these conclusions have been upheld with only minor exceptions. The triplet binding technique was a major innovation in the effort to decipher the genetic code.

Repeating Copolymers

Yet another innovative technique for deciphering the genetic code was developed in the early 1960s by Har Gobind Khorana, who was able to chemically synthesize long RNA molecules consisting of short ribonucleotide sequences repeated many times. First, he created the

TABLE 13.2 Amino Acid Assignments to Specific Trinucleotides Derived from the Triplet Binding Assay

Trinucleotides	Amino Acid
AAA AAG	Lysine
AUG	Methionine
AUU AUC AUA	Isoleucine
CCU CCC CCG CCA	Proline
CUC CUA CUG CUU	Leucine
GAA GAG	Glutamic acid
UCU UCC UCA UCG	Serine
UGU UGC	Cysteine
UUA UUG	Leucine
UUU UUC	Phenylalanine

individual short sequences (e.g., di-, tri-, and tetranucleotides); then he replicated them many times and finally joined them enzymatically to form the long polynucleotides, referred to as copolymers. In **Figure 13.6**, a dinucleotide made in this way is converted to an mRNA with two repeating triplet codons. A trinucleotide is converted into an mRNA that can be read as containing one of three potential repeating triplets, depending on the point at which initiation occurs. Finally, a tetranucleotide creates a message with four repeating triplet sequences.

When these synthetic mRNAs were added to a cell-free system, the predicted proportions of amino acids were found to be incorporated in the resulting polypeptides. When these data were combined with data drawn from mixed copolymer and triplet binding experiments, specific assignments were possible.

One example of specific assignments made in this way will illustrate the value of Khorana's approach. Consider the following experiments in concert with one another:

1. The repeating *trinucleotide sequence* UUCUUCUUC . . . can be read as three possible repeating triplets—UUC, UCU, and CUU—depending on the initiation point. When placed in a cell-free translation system, three different polypeptide homopolymers—containing phenylalanine, serine, or leucine—are produced. Thus, we know that each of the three triplets encodes one of the three amino acids, but we do not know which codes which.

2. On the other hand, the *repeating dinucleotide sequence* UCUCUCUC . . . produces the triplets UCU and CUC and, when used in an experiment, leads to the incorporation of leucine and serine into a polypeptide. Thus, the triplets UCU and CUC specify leucine and serine, but we still do not know which triplet specifies which amino acid. However, when considering both sets of results in concert, we can conclude that UCU, which is common to both experiments, must encode either leucine or serine but not phenylalanine. Thus, either CUU *or* UUC encodes leucine *or* serine, while the other encodes phenylalanine.

3. To derive more specific information, we can examine the results of using the repeating tetranucleotide sequence UUAC, which produces the triplets UUA, UAC, ACU, and CUU. The CUU triplet is one of the two in which we are interested. Three amino acids are incorporated by this experiment: leucine, threonine, and tyrosine. Because CUU must specify only serine or leucine, and because, of these two, only leucine appears in the resulting polypeptide, we may conclude that CUU specifies leucine. Once this assignment is established, we can logically determine all others. Of the two triplet pairs remaining (UUC and UCU from the first experiment *and* UCU and CUC from the second experiment), whichever triplet is common to both must encode serine. This is UCU. By elimination, UUC is determined to encode phenylalanine and CUC is determined to encode leucine.

Thus, through painstaking logical analysis, four specific triplets encoding three different amino acids have been

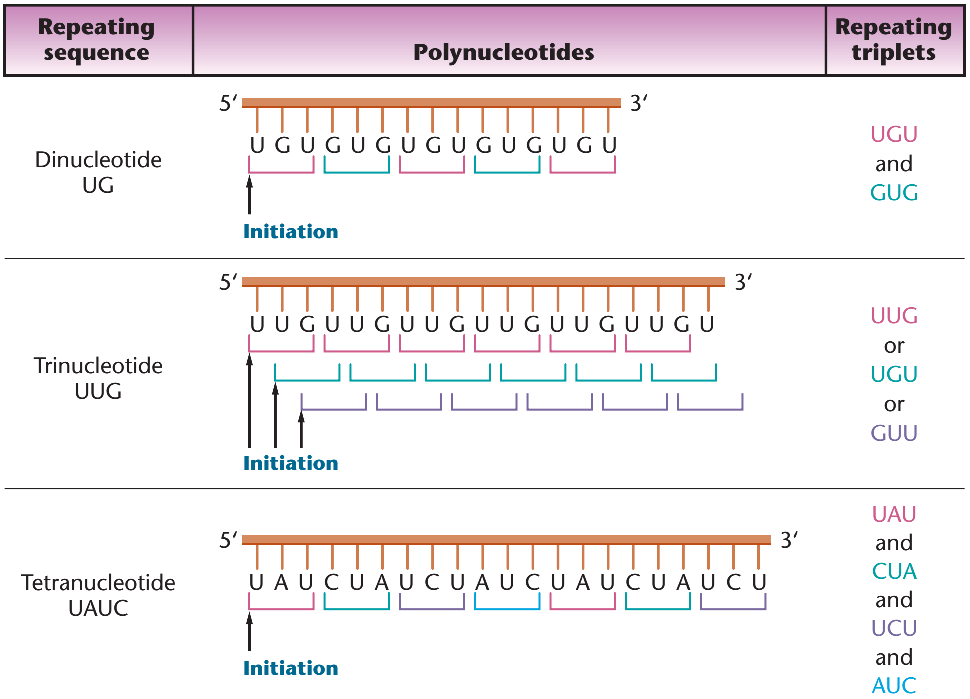


FIGURE 13.6 The conversion of di-, tri-, and tetranucleotides into repeating copolymers. The triplet codons produced in each case are shown.

TABLE 13.3 Amino Acids Incorporated Using Repeated Synthetic Copolymers of RNA

Repeating Copolymer	Codons Produced	Amino Acids in Resulting Polypeptides
UG	UGU	Cysteine
	GUG	Valine
AC	ACA	Threonine
	CAC	Histidine
UUC	UUC	Phenylalanine
	UCU	Serine
	CUU	Leucine
AUC	AUC	Isoleucine
	UCA	Serine
	CAU	Histidine
UAUC	UAU	Tyrosine
	CUA	Leucine
	UCU	Serine
	AUC	Isoleucine
GAUA	GAU	None
	AGA	None
	UAG	None
	AUA	None

assigned from these experiments. From these and similar interpretations, Khorana reaffirmed the identity of triplets that had already been deciphered and filled in gaps left from other approaches. Many examples are shown in **Table 13.3**.

The use of two tetranucleotide sequences, GAUA and GUAA, suggested that at least two triplets were *termination codons*. Khorana reached this conclusion because neither of these repeating sequences directed the incorporation of more than a few amino acids into a polypeptide, too few to detect. There are no triplets common to both messages, and both seemed to contain at least one triplet that terminates protein synthesis. Of the possible triplets in the poly-(GAUA) sequence, shown in Table 13.3, UAG was later shown to be a termination codon.

NOW SOLVE THIS

13.2 When repeating copolymers are used to form synthetic mRNAs, dinucleotides produce a single type of polypeptide that contains only two different amino acids. On the other hand, using a trinucleotide sequence produces three different polypeptides, each consisting of only a single amino acid. Why? What will be produced when a repeating tetranucleotide is used?

- **HINT:** *This problem asks you to consider different outcomes of repeating copolymer experiments. The key to its solution is to be aware that when using a repeating copolymer of RNA, translation can be initiated at different ribonucleotides. You must simply determine the number of triplet codons produced by initiation at each of the different ribonucleotides.*

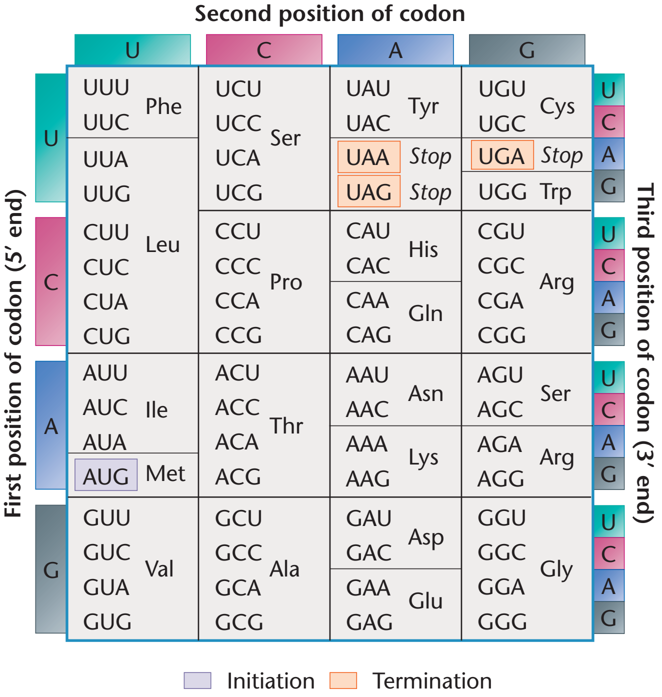
For more practice, see Problems 4–6.

13.4 The Coding Dictionary Reveals Several Interesting Patterns among the 64 Codons

The various techniques applied to decipher the genetic code have yielded a dictionary of 61 triplet codons assigned to amino acids. The remaining three codons are termination signals, not specifying any amino acid.

Degeneracy and the Wobble Hypothesis

A general pattern of triplet codon assignments becomes apparent when we look at the genetic coding dictionary. **Figure 13.7** displays the assignments in a particularly revealing form first suggested by Francis Crick.



First position of codon (5′ end)	Second position of codon: U		C		A		G		Third position of codon (3′ end)
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU		CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

▢ Initiation ▢ Termination

FIGURE 13.7 The genetic coding dictionary. AUG encodes methionine, which initiates most polypeptide chains. All other amino acids except tryptophan, which is encoded only by UGG, are represented by two to six triplets. The triplets UAA, UAG, and UGA are termination signals and do not encode any amino acids. Three-letter abbreviations for each amino acid are commonly used (see Figure 14.17) and are depicted here.

Most evident is that the code is degenerate, as the early researchers predicted. That is, almost all amino acids are specified by two, three, or four different codons. Three amino acids (serine, arginine, and leucine) are each encoded by six different codons. Only tryptophan and methionine are encoded by single codons.

Also evident is the *pattern* of degeneracy. Most often sets of codons specifying the same amino acid are grouped, such that the first two letters are the same, with only the third differing. For example, as you can see in Figure 13.7, the codons for phenylalanine (UUU and UUC in the top left corner of the coding table) differ only by their third letter. Either U or C in the third position specifies phenylalanine. Four codons specify valine (GUU, GUC, GUA, and GUG, in the bottom left corner), and they differ only by their third letter. In this case, all four letters in the third position specify valine.

Crick observed this pattern in the degeneracy throughout the code, and in 1966, he postulated the **wobble hypothesis.** Crick's hypothesis predicted that the initial two ribonucleotides of triplet codes are often more critical than the third member in attracting the correct tRNA. He postulated that hydrogen bonding at the third mRNA position of the codon–anticodon interaction would be *less* spatially constrained and need not adhere as strictly to the established base-pairing rules. The wobble hypothesis proposes a more flexible set of base-pairing rules at the third position of the codon (**Table 13.4**).

This relaxed base-pairing requirement, or "wobble," allows the anticodon of a single form of tRNA to pair with more than one triplet in mRNA. Consistent with the wobble hypothesis and the degeneracy of the code, U at the first position (the 5′ end) of the tRNA anticodon may pair with A or G at the third position (the 3′ end) of the mRNA codon, and G may likewise pair with U or C. Inosine (I), one of several modified bases found in tRNA (described in Chapter 14), may pair with C, U, or A. Based on these wobble rules, only about 30 different tRNA species are necessary to accommodate the 61 amino acid specifying codons. If nothing else, wobble can be considered an economy measure, assuming that the fidelity of translation is not compromised. We now know that, in actuality, there are 30 to 40 tRNA species present in bacteria and 41 to 55 tRNA species present in eukaryotes.

TABLE 13.4 Anticodon–Codon Base-Pairing Rules

Base at First Position (5′ End) of tRNA Anticodon	Base at Third Position (3′ End) of mRNA Codon
A	U
C	G
G	C or U
U	A or G
I	A, U, or C

The Ordered Nature of the Code

Still another observation has been made concerning the pattern of codon sequences and their corresponding amino acids, leading to the description referred to as the **ordered genetic code.** By this is meant that *chemically similar amino acids* often share one or two "middle" bases in the different triplets encoding them. For example, either U or C is often present in the second position of triplets that specify certain hydrophobic amino acids, including valine and alanine, among others. Two codons (AAA and AAG) specify the positively charged amino acid lysine. If only the middle letter of these codons is changed from A to G (AGA and AGG), the positively charged amino acid arginine is specified.

The chemical properties of amino acids will be discussed in more detail later in the text (see Chapter 14). For now, it is sufficient to note that the end result of an "ordered" code is that it buffers the potential effect of mutation on protein function. While many mutations of the second base of triplet codons result in a change of one amino acid to another, the change is often to an amino acid with similar chemical properties. In such cases, protein function may not be noticeably altered.

Punctuating the Code: Initiation and Termination Codons

In contrast to the *in vitro* experiments discussed earlier, initiation of protein synthesis *in vivo* is a highly specific process. The initial amino acid inserted into all polypeptide chains is methionine. (This residue is formylated in bacteria and unformylated in eukaryotes, as will be discussed in Chapter 14.) Only one codon, AUG, codes for methionine, and it is sometimes called the **initiator** (or **start**) **codon.** Rarely, two other codons, GUG and UUG, specify methionine during initiation, though it is not clear why this happens, since GUG normally encodes valine and UUG encodes leucine.

As mentioned in the preceding section, three other codons (UAG, UAA, and UGA) serve as **termination (or stop) codons,** punctuation signals that do not code for any amino acid. They are not recognized by a tRNA molecule, and translation terminates when they are encountered. Mutations that produce any of these three codons internally in a coding sequence also result in termination. In that case, only a partial polypeptide is synthesized, since it is prematurely released from the ribosome. When such a change occurs in the DNA, it is called a **nonsense mutation**.*

* Historically, the terms *amber* (UAG), *ochre* (UAA), and *opal* (UGA) were used to distinguish mutations producing any of the three termination codons.

13.5 The Genetic Code Has Been Confirmed in Studies of Phage MS2

The various aspects of the genetic code discussed so far yield a fairly complete picture. The code is triplet in nature, degenerate, unambiguous, and commaless, although it contains punctuation in the form of start and stop signals. These individual principles have been confirmed by the detailed analysis of the RNA-containing bacteriophage MS2 by Walter Fiers and his coworkers.

MS2 is a virus that infects the bacterium *E. coli*. Its genetic material is single-stranded RNA. Its genome—one of the smallest known—consists of only about 3500 ribonucleotides making up only four genes. These genes specify a coat protein, an RNA-directed replicase, a lysis protein, and a maturation protein (the A protein). This simple system of a small genome and few gene products allowed Fiers and his colleagues to sequence the genes and their products.

The amino acid sequence of the MS2 coat protein was completed in 1970, and the nucleotide sequence of the gene encoding the protein was reported in 1972. When the chemical compositions of this gene and its encoded protein are compared, they are found to exhibit *colinearity* with one another. That is, the linear sequence of triplet codons formed by the ribonucleotides corresponds precisely with the linear sequence of amino acids in the protein. Furthermore, the codon for the first amino acid is AUG, the common initiator codon; the codon for the last amino acid is followed by two consecutive termination codons, UAA and UAG. We will return to discuss this concept later in the text (see Chapter 14).

By 1976, two other genes of MS2 and their protein products were sequenced. The analysis clearly showed that the genetic code in this virus was identical to that which had been established in bacterial systems. Other evidence suggests that the code is also identical in eukaryotes, thus providing confirmation of what seemed to be a universal code.

13.6 The Genetic Code Is *Nearly* Universal

Between 1960 and 1978, it was generally assumed that the genetic code would be found to be universal, applying equally to viruses, bacteria, archaea, and eukaryotes. Certainly, the nature of mRNA and the translation machinery seemed to be very similar in these organisms. For example, cell-free systems derived from bacteria could translate eukaryotic mRNAs. Poly U stimulates synthesis of polyphenylalanine in cell-free systems when the components are derived from eukaryotes. Early studies involving recombinant DNA technology (Chapter 20) revealed that eukaryotic genes can be inserted into bacterial cells, where they are then transcribed and translated. Within eukaryotes, mRNAs from mice and rabbits were injected into amphibian eggs and efficiently translated. For the many eukaryotic genes that had been sequenced at the time, notably those for hemoglobin molecules, the amino acid sequence of the encoded proteins adhered to the coding dictionary established from bacterial studies.

However, several 1979 reports on the coding properties of DNA derived from mitochondria (mtDNA) from yeast and humans began to undermine the hypothesis of the universality of the genetic language. Cloned mtDNA fragments were sequenced and compared with the amino acid sequences of various mitochondrial proteins, revealing several exceptions to the coding dictionary (**Table 13.5**). Most surprising was that the codon UGA, normally specifying termination, specifies the insertion of tryptophan during translation in yeast and human mitochondria. In human mitochondria, AUA, which normally specifies isoleucine, directs the internal insertion of methionine. In yeast mitochondria, threonine is inserted instead of leucine when CUA is encountered in mRNA.

In 1985, exceptions to the standard coding dictionary were also discovered in the bacterium *Mycoplasma capricolum*, and in the nuclear genes of the protozoan ciliates *Paramecium, Tetrahymena,* and *Stylonychia*. For example, as shown in Table 13.5, one alteration converts the termination codon (UGA) to tryptophan. Several other code alterations convert termination codons (UAA, UAG) to glutamine. These changes are significant because they are seen in both a bacterium and several eukaryotes, that is, in distinct species that have evolved separately over a long period of time.

Note the pattern apparent in several of the altered codon assignments: The change in coding capacity involves only a shift in recognition of the third, or wobble, position.

TABLE 13.5 Exceptions to the Universal Code

Codon	Normal Code Word	Altered Code Word	Source
UGA	Termination	Trp	Human and yeast mitochondria; *Mycoplasma*
CUA	Leu	Thr	Yeast mitochondria
AUA	Ile	Met	Human mitochondria
AGA	Arg	Termination	Human mitochondria
AGG	Arg	Termination	Human mitochondria
UAA	Termination	Gln	*Paramecium, Tetrahymena,* and *Stylonychia*
UAG	Termination	Gln	*Paramecium*

For example, AUA specifies isoleucine in the cytoplasm and methionine in the mitochondrion, but in the cytoplasm, methionine is specified by AUG. Similarly, UGA calls for termination in the cytoplasm, but for tryptophan in the mitochondrion; in the cytoplasm, tryptophan is specified by UGG. It has been suggested that such changes in codon recognition may represent an evolutionary trend toward reducing the number of tRNAs needed in mitochondria; only 22 tRNA species are encoded in human mitochondria, for example. However, until other cases are discovered, the differences must be considered as exceptions to the previously established general coding rules.

13.7 Different Initiation Points Create Overlapping Genes

Earlier we stated that the genetic code is nonoverlapping, meaning that each ribonucleotide in the code *for a given polypeptide* is part of only one codon. However, this characteristic of the code does not rule out the possibility that a single mRNA may have multiple initiation points for translation. If so, these points could theoretically create several different reading frames within the same mRNA, thus specifying more than one polypeptide. This concept, which creates **overlapping genes,** is illustrated in Figure 13.8(a).

Note that a gene is also referred to as an **ORF,** or **open reading frame,** which is defined as any DNA sequence that produces a functional mRNA, one with a start and stop codon between which is a series of triplet codons specifying the amino acids making up a polypeptide. Thus, we sometimes also refer to *overlapping ORFs*.

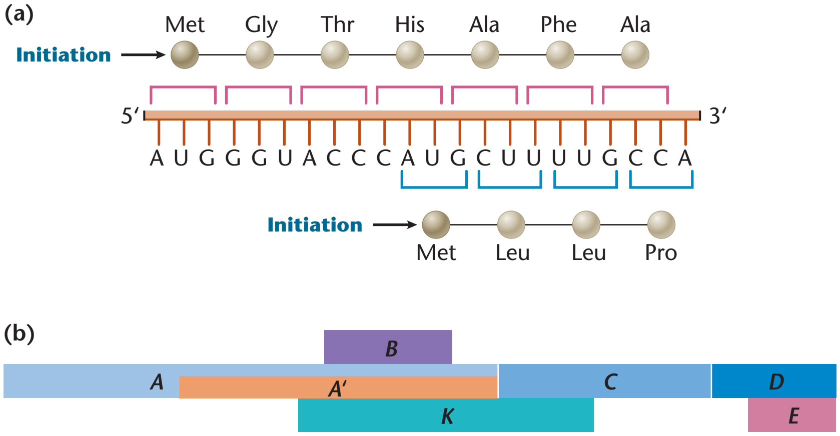


FIGURE 13.8 Illustration of the concept of overlapping reading frames. (a) Translation initiated at two different AUG positions out of frame with one another will give rise to two distinct amino acid sequences. (b) The relative positions of the sequences encoding seven polypeptides of the phage ϕX174.

That such overlapping might actually occur in some viruses was suspected when phage ϕX174 was carefully investigated. The circular chromosome (the first DNA-based genome to be fully sequenced) consists of 5386 nucleotides, which should encode a maximum of 1795 amino acids, sufficient for five or six average-sized proteins. However, this small virus in fact synthesizes 11 proteins consisting of more than 2300 amino acids. A comparison of the nucleotide sequence of the DNA and the amino acid sequences of the polypeptides synthesized has clarified the apparent paradox. At least four instances of multiple initiation have been discovered, creating overlapping genes [Figure 13.8(b)].

The sequences specifying the K and B polypeptides are initiated in separate reading frames within the sequence specifying the A polypeptide. The *K* gene sequence overlaps into the adjacent sequence specifying the C polypeptide. The *E* sequence is out of frame with, but initiated within, that of the D polypeptide. Finally, the *A′* sequence, while in frame with the *A* sequence, is initiated in the middle of the *A* sequence. They both terminate at the identical point. In all, seven different polypeptides are created from a DNA sequence that might otherwise have specified only three (A, C, and D).

A similar situation has been observed in other viruses and bacteria. The employment of overlapping reading frames optimizes the limited amount of genetic material present. However, such an approach to storing information has a distinct disadvantage in that a single mutation may affect more than one protein and thus increase the chances that the change will be deleterious or lethal. In the example we just discussed [Figure 13.8(b)], a single mutation in the middle of the *B* gene could potentially affect three other proteins (the A, A′, and K proteins). It may be for this reason that, while present, overlapping genes are not common in other organisms.

13.8 Transcription Synthesizes RNA on a DNA Template

Even while the genetic code was being studied, it was quite clear that proteins were the end products of many genes. Hence, while some geneticists were attempting to elucidate the code, other research efforts were directed toward the nature of genetic expression. The central question was how DNA, a nucleic acid, can specify a protein composed of amino acids.

The complex, multistep process begins with the transfer of genetic information stored in DNA to RNA. The process by which RNA molecules are synthesized on a DNA template is called **transcription.** It results in an mRNA molecule complementary to the gene sequence of one of the two strands of the double helix. Each triplet codon in the mRNA is, in turn, complementary to the anticodon region of its corresponding tRNA, which inserts the correct amino acid into the polypeptide chain during translation.

The significance of transcription is enormous, for it is the initial step in the process of *information flow* within the cell. The idea that RNA is involved as an intermediate molecule in the process of information flow between DNA and protein was suggested by the following findings:

1. DNA is, for the most part, associated with chromosomes in the nucleus of the eukaryotic cell. However, protein synthesis occurs in association with ribosomes located outside the nucleus, in the cytoplasm. Therefore, DNA did not appear to participate directly in protein synthesis.
2. RNA is synthesized in the nucleus of eukaryotic cells, in which DNA is found, and is chemically similar to DNA.
3. Following its synthesis, most messenger RNA migrates to the cytoplasm, in which protein synthesis (translation) occurs.

Collectively, these observations suggested that genetic information, stored in DNA, is transferred to an RNA intermediate, which directs the synthesis of the proteins. As with most new ideas in molecular genetics, the initial supporting experimental evidence for an RNA intermediate was based on studies of bacteria and bacteriophages. The results of these experiments agreed with the concept of a messenger RNA (mRNA) being made on a DNA template and then directing the synthesis of specific proteins in association with ribosomes. This concept was formally proposed by François Jacob and Jacques Monod in 1961 as part of a model for gene regulation in bacteria. Since then, mRNA has been isolated and studied thoroughly. There is no longer any question about its role in genetic processes.

13.9 RNA Polymerase Directs RNA Synthesis

To prove that RNA can be synthesized on a DNA template, it was necessary to demonstrate that there is an enzyme capable of directing this synthesis. By 1959, several investigators, including Samuel Weiss, had independently discovered such a molecule in rat liver. Called **RNA polymerase,** it has the same general substrate requirements as does DNA polymerase, the major exception being that the substrate nucleotides contain the ribose rather than the deoxyribose form of the sugar. Unlike DNA polymerase, no primer is required to initiate synthesis. The overall reaction summarizing the synthesis of RNA on a DNA template can be expressed as

$$n(\text{NTP}) \xrightarrow{\substack{\text{RNA} \\ \text{polymerase}}} (\text{NMP})_n + n(\text{PP}_i)$$

As the equation reveals, nucleoside triphosphates (NTPs) serve as substrates for the enzyme, which catalyzes the polymerization of nucleoside monophosphates (NMPs), or nucleotides, into a polynucleotide chain $(\text{NMP})_n$. Nucleotides are linked during synthesis by 5′ to 3′ phosphodiester bonds (see Figure 10.12). The energy released by cleaving the triphosphate precursor into the monophosphate form drives the reaction, and inorganic diphosphates (PP_i) are produced.

A second equation summarizes the sequential addition of each ribonucleotide as the process of transcription progresses:

$$(\text{NMP})_n + \text{NTP} \xrightarrow{\substack{\text{RNA} \\ \text{polymerase}}} (\text{NMP})_{n+1} + \text{PP}_\text{i}$$

As this equation shows, each step of transcription involves the addition of one ribonucleotide (NMP) to the growing polyribonucleotide chain $(\text{NMP})_{n+1}$, using a nucleoside triphosphate (NTP) as the precursor.

RNA polymerase from *E. coli* has been extensively characterized. The **core enzyme** has been shown to consist of subunits designated α (two copies), β, β', and ω. A slightly more complex form of the enzyme, the **holoenzyme,** contains the additional subunit σ and has a molecular weight of almost 500 kilodaltons (kDa). While there is some variation in the subunit composition of other bacteria, it is the β and β' subunits that provide the catalytic mechanism and active site for transcription. As we will see, the **sigma (σ) factor** [**Figure 13.9(a)**] plays a regulatory function in the initiation of RNA transcription.

While there is but a single form of the core enzyme in *E. coli,* there are several different σ factors, creating variations of the polymerase holoenzyme. On the other hand, eukaryotes display three distinct forms of RNA polymerase, each consisting of a greater number of polypeptide subunits than in bacteria. In this section, we will discuss the process of transcription in bacteria. We will return to a discussion of eukaryotic transcription later in this chapter.

Promoters, Template Binding, and the σ Subunit

Transcription results in the synthesis of a single-stranded RNA molecule complementary to a region along only one of the two strands of the DNA double helix. When discussing

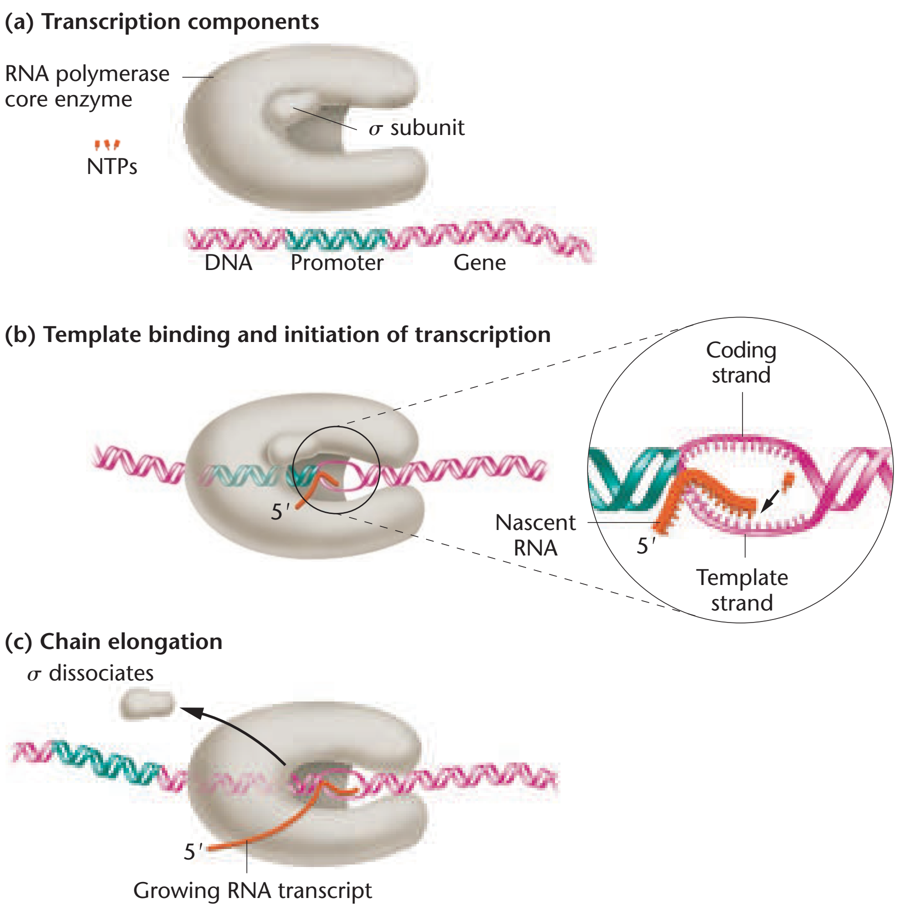


FIGURE 13.9 The early stages of transcription in bacteria, showing (a) the components of the process; (b) template binding at the –10 site involving the sigma subunit of RNA polymerase and subsequent initiation of RNA synthesis; and (c) chain elongation, after the σ subunit has dissociated from the transcription complex and the enzyme moves along the DNA template.

transcription, the DNA strand that serves as a template for RNA polymerase is denoted as the **template strand** and the complementary DNA strand is called the **coding strand.** Note that the complementary strand is called the coding strand because it and the RNA molecule transcribed from the template strand have the same 5′ to 3′ nucleotide sequence, but with uridine (U) substituted for thymidine (T) in the RNA.

The initial step in bacterial gene transcription is referred to as **template binding** [Figure 13.9(b)]. The site of this initial binding is established when the RNA polymerase σ subunit recognizes specific DNA sequences called **promoters.** These recognition sequences are located in the 5′ region, upstream from the coding sequence of a gene. It is believed that the holoenzyme "explores" a length of DNA until it encounters a promoter region and binds there to about 60 nucleotide pairs along the double helix, 40 of which are upstream from the point of initial transcription. Once this occurs, the helix is denatured, or unwound, locally, making the template strand of the DNA accessible to the enzyme. The point at which transcription begins is called the **transcription start site,** often indicated as position +1.

Because the interaction of promoters with RNA polymerase governs the efficiency of transcription—by regulating the initiation of transcription—the importance of promoter sequences cannot be overemphasized. The nature of the binding between polymerase and promoters is at the heart of future discussions concerning genetic regulation, the subject of later chapters in the text (see Chapters 16, 17, and 18). While those chapters present more detailed information concerning enzyme–promoter interactions, we address three points here.

The first point is the concept of **consensus sequences,** DNA sequences that are similar (homologous) in different genes of the same organism or in one or more genes of related organisms. Their conservation during evolution attests to the critical nature of their role in biological processes. Two consensus sequences have been found in bacterial promoters. One, TATAAT, is located 10 nucleotides upstream from the site of initial transcription (the −10 region, or **Pribnow box**). The other, TTGACA, is located 35 nucleotides upstream (the −35 region). Mutations in either region diminish transcription, often severely.

DNA sequences such as these, located near the region to be transcribed, are said to be ***cis*-acting elements.** The term *cis,* drawn from organic chemistry nomenclature, means "next to" or on the same side as other functional groups, in contrast to being *trans* to or "across from" them. In molecular genetics, then, *cis*-elements are regulatory sequences located within the same molecule of DNA as the target they regulate. In contrast, ***trans*-acting factors** are proteins (e.g., σ factor) that influence gene expression by binding to *cis*-acting elements.

The second point is that the degree of RNA polymerase binding to different promoters varies greatly, causing variable gene expression. Currently, this is attributed to sequence variation in the promoters. In bacteria, both strong promoters and weak promoters have been discovered, causing a variation in time of initiation from once every 1 to 2 seconds to as little as once every 10 to 20 minutes. Mutations in promoter sequences may severely reduce the initiation of gene expression.

The third point about polymerase—promoter interactions involves the σ subunit in bacteria. In *E. coli*, for example, the major form is designated σ^{70} based on its molecular weight of 70 kDa. σ^{70} recognizes the promoters of most genes in the cell. However, several other σ factors are also present (e.g., σ^{32}, σ^{54}, σ^{S}, and σ^{E}), which are called upon to regulate gene expression under different environmental conditions such as heat or starvation. Each σ factor recognizes different promoter sequences, which in turn provides specificity to the initiation of transcription.

NOW SOLVE THIS

13.3 The following represent deoxyribonucleotide sequences in the template strand of DNA:

Sequence 1:	5′-CTTTTTTGCCAT-3′
Sequence 2:	5′-ACATCAATAACT-3′
Sequence 3:	5′-TACAAGGGTTCT-3′

(a) For each strand, determine the mRNA sequence that would be derived from transcription.

(b) Using Figure 13.7, determine the amino acid sequence that is encoded by these mRNAs.

(c) For sequence 1, what is the sequence of the coding DNA strand?

■ **HINT:** *This problem asks you to consider the outcome of the transfer of complementary information from DNA to RNA and to determine the amino acids encoded by this information. The key to its solution is to remember that in RNA, uracil is complementary to adenine, and that while DNA stores genetic information in the cell, the code that is ultimately translated is contained in the RNA and is complementary to the template strand of DNA.*

Initiation, Elongation, and Termination of RNA Synthesis in Bacteria

Once RNA polymerase has recognized and bound to the promoter, DNA is locally converted from its double-stranded form to an open structure, exposing the template strand. The enzyme then proceeds to initiate RNA synthesis, whereby the first ribonucleoside triphosphate, which is complementary to the first template nucleotide, is inserted at the start site. This will be the 5′ end of the transcript. (As we noted earlier, unlike in DNA synthesis, no primer is required.) Subsequent ribonucleotide complements are inserted and linked together by phosphodiester bonds as RNA polymerization proceeds (in a 5′ to 3′ direction in terms of the nascent RNA). This process continues, creating a temporary 8-bp DNA/RNA duplex whose chains run antiparallel to one another [Figure 13.9(b), inset].

After these initial ribonucleotides have been added to the growing RNA chain, the σ subunit dissociates from the holoenzyme, and **chain elongation** proceeds under the direction of the core enzyme [Figure 13.9(c)]. In *E. coli*, elongation proceeds at the rate of about 50 nucleotides/second at 37°C.

Like DNA polymerase, RNA polymerase can perform **proofreading** as it adds each nucleotide. Proofreading leads to the recognition of mismatches where a noncomplementary base has been inserted. In such a case, the enzyme backs up and removes the mismatch. It then reverses direction and continues elongation. The RNA molecule being synthesized will be precisely complementary to the DNA sequence of the template strand of the gene. Wherever an A, T, C, or G residue is encountered, a corresponding U, A, G, or C residue, respectively, is incorporated into the RNA molecule. Transcriptional fidelity is important since RNA molecules ultimately provide the information leading to the synthesis of all proteins in the cell.

The enzyme traverses the entire gene until eventually it encounters a specific nucleotide sequence that acts as a termination signal. Such termination sequences are extremely important in bacteria because of the close proximity between the end of one gene and the upstream sequences of the adjacent gene.

An interesting aspect of termination in bacteria is that the termination sequence alluded to above is actually transcribed into RNA. The unique sequence of ribonucleotides in this termination region causes the newly formed transcript to fold back on itself, forming what is called a **hairpin secondary structure,** held together by hydrogen bonds. There are two different types of transcription termination mechanisms in bacteria, both of which are dependent on the formation of a hairpin structure in the RNA being transcribed.

Roughly 80 percent of transcripts in *E. coli* are terminated by **intrinsic termination** [Figure 13.10(a)]. Intrinsic termination is dependent on self-complementary GC-rich sequences (inverted repeats) within the transcript, which form a stable GC-rich hairpin structure, immediately followed by a string of uracil residues. The GC-rich hairpin structure causes RNA polymerase to stall during transcription of the adjacent poly U tract. The U bases within this region of the transcript have a relatively weak interaction with the A bases on the template strand of the DNA because there are only two hydrogen bonds per base pair. This leads to dissociation of RNA polymerase and the transcript is released.

The other well-described bacterial transcription termination mechanism is known as **rho-dependent termination** [Figure 13.10(b)], which is used for roughly 20 percent of the genes in *E. coli*. This mechanism is dependent on the **termination factor, rho (ρ)** and a termination sequence that is transcribed into a hairpin structure in the transcript. Rho is a large hexameric

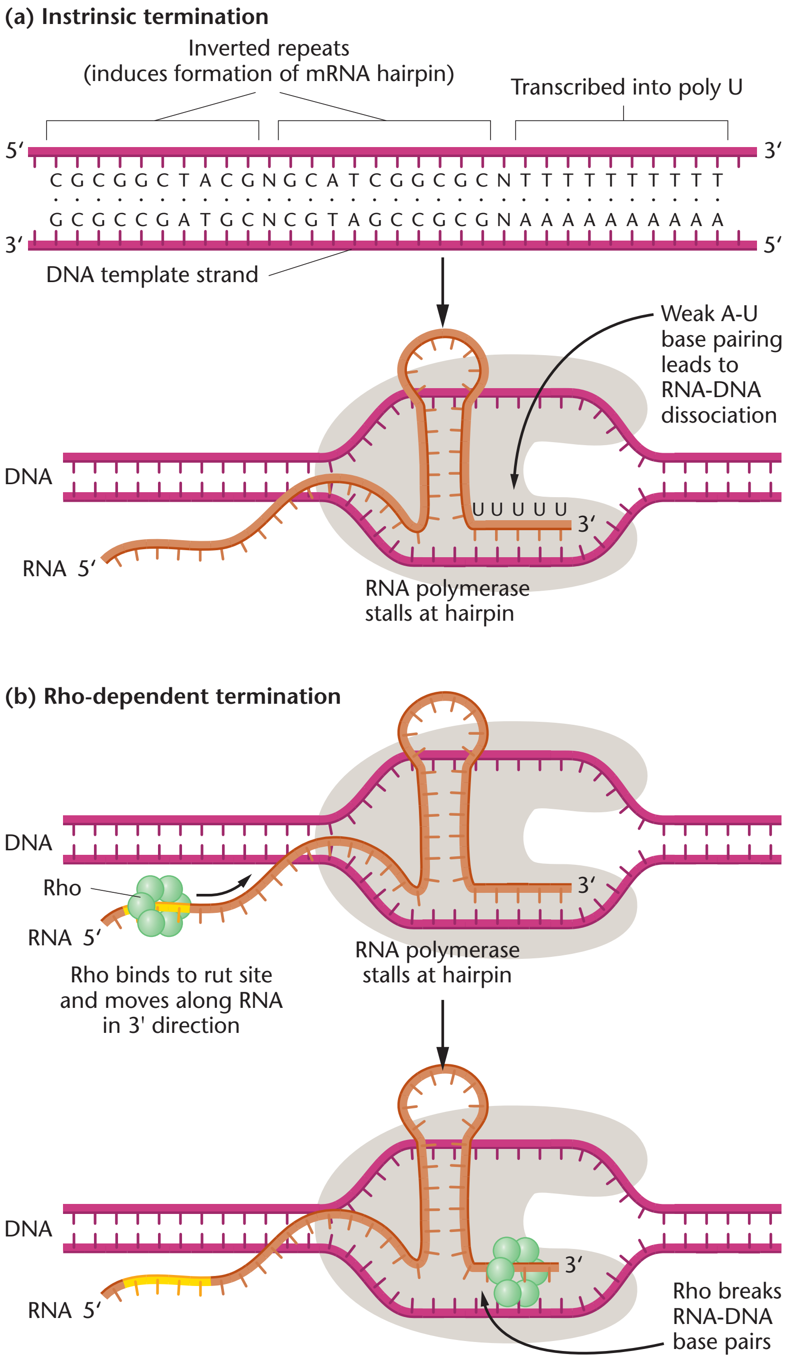


FIGURE 13.10 Transcription termination in bacteria. Intrinsic termination (a) involves a hairpin structure followed by a string of repeated U residues. Rho-dependent termination (b) involves the termination factor rho and a hairpin structure.

protein with RNA helicase activity—it can dissociate RNA secondary structures such as hairpins and DNA/RNA interactions. Rho binds to a specific sequence on the transcript known as the **rho utilization site (rut)** as soon as it is transcribed. Rho then moves along the transcript toward the 3′ end chasing after RNA polymerase. When RNA polymerase reaches the hairpin structure encoded by the termination sequence, it pauses and Rho catches up. Rho moves through the hairpin with its RNA helicase activity and then causes dissociation of RNA polymerase by breaking the hydrogen bonds between the DNA template and the transcript.

13.10 Transcription in Eukaryotes Differs from Bacterial Transcription in Several Ways

Much of our knowledge of transcription has been derived from studies of bacteria. Most of the general aspects of the mechanics of these processes are similar in eukaryotes, but there are several notable differences:

1. Transcription in eukaryotes occurs within the nucleus. Thus, unlike the bacterial process, in eukaryotes the

RNA transcript is not free to associate with ribosomes prior to the completion of transcription. For the mRNA to be translated, it must move out of the nucleus into the cytoplasm.

2. Transcription in eukaryotes occurs under the direction of *three separate forms* of RNA polymerase, rather than the single form seen in bacteria.

3. Initiation of transcription of eukaryotic genes requires that compact chromatin fiber, characterized by nucleosome coiling, be uncoiled to make the DNA helix accessible to RNA polymerase and other regulatory proteins. This transition, referred to as *chromatin remodeling* (first introduced in Chapter 12 and discussed in more detail in Chapters 17 and 19), reflects the dynamics involved in the conformational change that occurs as the DNA helix is opened.

4. Initiation and regulation of transcription entail a more extensive interaction between *cis*-acting DNA sequences and *trans*-acting protein factors. For example, while bacterial RNA polymerase requires only a σ subunit to bind the promoter and initiate transcription, in eukaryotes, several *general transcription factors* (*GTFs*) are required to bind the promoter, recruit RNA polymerase, and initiate transcription. Furthermore, in addition to promoters, eukaryotic genes often have other *cis*-acting control units called *enhancers* and *silencers* (discussed below, and in more detail in Chapter 17), which greatly influence transcriptional activity.

5. In bacteria, transcription termination is often dependent upon the formation of a hairpin secondary structure in the transcript. However, eukaryotic transcription termination is more complex. Transcriptional termination for protein-coding genes involves sequence-specific cleavage of the transcript, which then leads to eventual dissociation of RNA polymerase from the DNA template.

6. In eukaryotes, the initial (or primary) transcripts of protein-coding mRNAs, called **pre-mRNAs,** undergo complex alterations, generally referred to as "processing," to produce a mature mRNA. Processing often involves the addition of a 5′ cap and a 3′ tail, and the removal of intervening sequences that are not a part of the mature mRNA. In the remainder of this chapter we will look at the basic details of transcription and mRNA processing in eukaryotic cells. The process of transcription is highly regulated, determining which DNA sequences are copied into RNA and when and how frequently they are transcribed. We will return to topics directly related to the regulation of eukaryotic gene transcription later in the text (see Chapter 17).

Initiation of Transcription in Eukaryotes

As noted earlier, eukaryotic RNA polymerase exists in three distinct forms. Each eukaryotic RNA polymerase is larger and more complex than the single form of RNA polymerase found in bacteria. For example, yeast and human RNA polymerase II enzymes consist of 12 subunits. While the three forms of the enzyme share certain protein subunits, each nevertheless transcribes different types of genes, as indicated in **Table 13.6**.

RNA polymerases I and III (RNAP I and RNAP III) transcribe transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), which are needed in essentially all cells at all times for the basic process of protein synthesis. In contrast, **RNA polymerase II (RNAP II),** which transcribes protein-coding genes, is highly regulated. Protein-coding genes are often expressed at different times, in response to different signals, and in different cell types. Thus, RNAP II activity is tightly regulated on a gene-by-gene basis. For this reason, most studies of transcription in eukaryotes have focused on RNAP II.

The activity of RNAP II is dependent on both the *cis*-acting regulatory elements of the gene and a number of *trans*-acting transcription factors that bind to these DNA elements. (We will consider *cis* elements first and then turn to *trans* factors.)

At least four different types of *cis*-acting DNA elements regulate the initiation of transcription by RNAP II. The first of these, the **core promoter,** includes the transcription start site. It determines where RNAP II binds to the DNA and where it begins transcribing the DNA into RNA. Another promoter element, called a **proximal-promoter element,** is located upstream of the start site and helps modulate the level of transcription. The last two types of *cis*-acting elements, called **enhancers,** and **silencers,** influence the efficiency or the rate of transcription initiation by RNAP II from the core-promoter element.

In some eukaryotic genes, a *cis*-acting element within the core promoter is the **Goldberg–Hogness box,** or **TATA box.** Located about 30 nucleotide pairs upstream (−30) from the start point of transcription, TATA boxes share a consensus sequence TATA$^{A}/_{T}$AAR, where R indicates any purine nucleotide. The sequence and function of TATA boxes are analogous to those found in the −10

TABLE 13.6 RNA Polymerases in Eukaryotes

Form	Product	Location
I	rRNA	Nucleolus
II*	mRNA, snRNA	Nucleoplasm
III	5*S*rRNA, tRNA	Nucleoplasm

*RNAP II also synthesizes a variety of other RNAs, including miRNAs and lncRNAs (see Chapters 18 and 19).

promoter region of bacterial genes. However, recall that in bacteria, RNA polymerase binds directly to the −10 promoter region. As we will see below, the same is not the case in eukaryotes.

Although eukaryotic promoter elements can determine the site and general efficiency of initiation, other elements—known as *enhancers* and *silencers*—have more dramatic effects on eukaryotic gene transcription. As their names suggest, enhancers increase transcription levels and silencers decrease them. The locations of these elements can vary from immediately upstream of a promoter to downstream, within, or kilobases away, from a gene. In other words, they can modulate transcription from a distance. Each eukaryotic gene has its own unique arrangement of promoter, enhancer, and silencer elements.

Complementing the *cis*-acting regulatory sequences are various *trans*-acting factors that facilitate RNAP II binding and, therefore, the initiation of transcription. These proteins are referred to as **transcription factors.** There are two broad categories of transcription factors: the **general transcription factors (GTFs)** that are absolutely required for all RNAP II–mediated transcription, and the **transcriptional activators** and **transcriptional repressors** that influence the efficiency or the rate of RNAP II transcription initiation.

The general transcription factors are essential because RNAP II cannot bind directly to eukaryotic core-promoter sites and initiate transcription without their presence. The general transcription factors involved with human RNAP II binding are well characterized and are designated **TFIIA, TFIIB,** and so on. One of these, **TFIID,** binds directly to the TATA-box sequence. Once initial binding of TFIID to DNA occurs, the other general transcription factors, along with RNAP II, bind sequentially to TFIID, forming an extensive **pre-initiation complex**.

Transcriptional activators and repressors bind to enhancer and silencer elements and regulate transcription initiation by aiding or preventing the assembly of pre-initiation complexes and the release of RNAP II from pre-initiation into full transcription elongation. They appear to supplant the role of the σ factor seen in the bacterial enzyme and are important in eukaryotic gene regulation. We will consider the roles of general and specific transcription factors in eukaryotic gene regulation, as well as the various DNA elements to which they bind, in more detail later in the text (Chapter 17).

Recent Discoveries Concerning Eukaryotic RNA Polymerase Function

It is of great interest to learn how RNA polymerase II is able to achieve transcription along the chromatin fiber. The enzyme must open up the DNA helix and locally separate (denature) the two strands so that the template strand may pass through its active site during RNA synthesis. The ability to crystallize large nucleic acid–protein structures and perform X-ray diffraction analysis at resolutions below 5 Å has shed light on this issue—particularly the work of Roger Kornberg and colleagues, using RNAP II isolated from yeast. (It is useful to note here that achieving a resolution below 2.8 Å allows the visualization of each amino acid of every protein in the complex!) Various discoveries have provided a highly detailed account of the most critical processes of transcription.

RNAP II in yeast contains two large protein subunits and ten smaller ones, forming a huge three-dimensional complex with a molecular weight of about 500 kDa. The promoter DNA is initially positioned over a cleft formed between the two large subunits of RNAP II. The subunit assemblage in fact resembles a pair of jaws that can open and partially close and at this stage is combined with the general transcription factor TFIIB. Prior to association with DNA, the cleft is open; once associated with DNA, the cleft partially closes, securing the duplex during the initiation of transcription. The part of the enzyme that is critical for this transition is about 50 kDa in size and is called the *clamp*.

Once secured by the clamp, the two strands of a small region of duplex DNA separate at a position within the enzyme referred to as the *active center,* and the template strand is scanned for the transcription start site. Complementary RNA synthesis is then initiated on the DNA template strand. However, the entire complex remains unstable, and transcription usually terminates following the incorporation of only a few ribonucleotides. It is not clear why, but this so-called *abortive transcription* is repeated a number of times before a stable DNA:RNA hybrid containing a transcript of 11 ribonucleotides is formed. Once this occurs, abortive transcription is overcome, a stable complex is achieved, and elongation of the RNA transcript proceeds in earnest. Transcription at this point is said to have achieved a level of highly processive RNA polymerization—meaning that the enzyme is able to catalyze numerous consecutive polymerization reactions without releasing its substrate.

As transcription proceeds, the enzyme moves along the DNA, and at any given time, about 40 base pairs of DNA and 18 residues of the growing RNA chain are part of the enzyme–substrate complex. The earliest synthesized RNA runs through a groove in the enzyme and exits under a structure at the top and back designated as the *lid*. Another area, called the *pore,* has been identified at the bottom of the enzyme. The pore serves as the point through which ribonucleoside triphosphates (NTPs), the RNA precursors, gain entry into the complex.

Unlike in bacteria, there is no specific sequence that signals for the termination of transcription. In fact, RNAP II often continues transcription well beyond what will be the eventual 3′ end of the mature mRNA. Once transcription has incorporated a specific sequence AAUAAA, known as the **polyadenylation signal sequence** (discussed below), the transcript is enzymatically cleaved roughly 10–35 bases further downstream in the 3′ direction. Cleavage of the transcript destabilizes RNAP II, much as it was during the earlier state of abortive transcription. The clamp opens, and both DNA and RNA are released from the enzyme as transcription is terminated. This completes the cycle that constitutes transcription.

In sum, an unstable DNA–enzyme complex is formed during the initiation of transcription, stability is established once elongation manages to create a DNA/RNA duplex of sufficient size, elongation proceeds, and then enzyme instability again characterizes termination of transcription. As you think back on this cycle, try to visualize the process mentally, from the time the DNA first associates with the enzyme until the transcript is released from the large molecular complex. If these images are clear to you, then you no doubt have acquired a firm understanding of transcription in eukaryotes, which is more complex than in bacteria.

Kornberg's findings have extended our knowledge of transcription considerably. For this work, he was awarded the Nobel Prize in Chemistry in 2006.

Processing Eukaryotic RNA: Caps and Tails

While in bacteria the base sequence of DNA is transcribed into an mRNA that is immediately and directly translated into the amino acid sequence as dictated by the genetic code, eukaryotic mRNAs require significant alteration before they are transported to the cytoplasm and translated. By 1970, evidence accumulated by James Darnell and others showed that eukaryotic mRNA is transcribed initially as a precursor molecule much larger than that which is translated into protein. It was proposed that this *primary transcript* of a gene (a *pre-mRNA*) must be processed in the nucleus before it appears in the cytoplasm as a *mature mRNA* molecule. The various processing steps, discussed in the sections that follow, are summarized in **Figure 13.11**.

An important **posttranscriptional modification** of eukaryotic RNA transcripts destined to become mRNAs occurs at the 5′ end of these molecules, where a **7-methylguanosine (m^7G) cap** is added. This cap, discovered by Aaron Shatkin and James Darnell, is added shortly after synthesis of the initial RNA transcript has begun and appears to be important to subsequent processing within the nucleus. The cap stabilizes the mRNA by protecting the 5′ end of the molecule from nuclease attack. Subsequently, the cap facilitates the transport of mature mRNAs from the nucleus into the cytoplasm, and is required for the initiation of translation of the mRNA into protein. Chemically, the cap is a guanosine residue with a methyl group (CH_3) at position 7 of the base. The cap is also distinguished by a unique 5′-to-5′ triphosphate bridge that connects it to the initial ribonucleotide of the RNA. Some eukaryotes also acquire a methyl group at the 2′-carbons of the ribose sugars of the first two ribonucleotides of the RNA.

Further insights into the processing of RNA transcripts during the maturation of mRNA came from the discovery that mRNAs contain, at their 3′ end, a stretch of as many as 250 adenylic acid residues. As discussed earlier in the context of eukaryotic transcription termination, the transcript is cleaved roughly 10–35 ribonucleotides after the highly conserved AAUAAA polyadenylation signal sequence. An enzyme known as **poly-A polymerase** then catalyzes the addition of a poly-A tail to the free 3′-OH group at the end of the transcript. Poly-A tails are found at the 3′ end of almost all mRNAs studied in a variety of eukaryotic organisms. The exceptions in eukaryotes seem to be mRNAs that encode histone proteins.

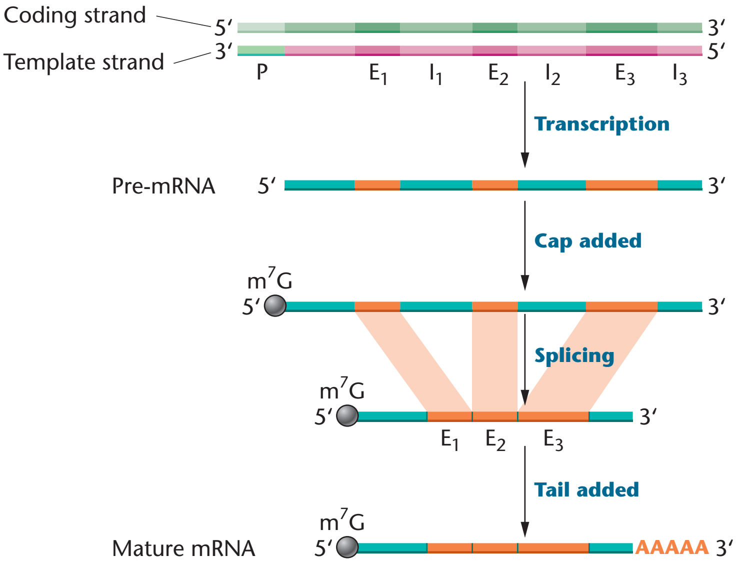


FIGURE 13.11 Posttranscriptional RNA processing in eukaryotes. Beginning at the promoter (P) of a gene, transcription produces a pre-mRNA containing several introns (I) and exons (E), as identified under the DNA template strand. Shortly after transcription begins, a m^7G cap is added to the 5′ end. Next, and during transcription elongation, the introns are spliced out and the exons joined. Finally, a poly-A tail is added to the 3′ end. While this figure depicts these steps sequentially, in some eukaryotic transcripts, the poly-A tail is added before splicing of all introns has been completed.

While the AAUAAA signal sequence is not found on all eukaryotic mRNAs, it appears to be essential to those that have it. If the sequence is changed as a result of a mutation, those transcripts that would normally have it cannot add the poly-A tail. In the absence of this tail, these RNA transcripts are rapidly degraded by nucleases. The **poly A binding protein** (see Chapter 18), as the name suggests, binds to poly-A tails and prevents nucleases from degrading the 3′ end of the mRNA. In addition, we know now that the poly-A tail is important for export of the mRNA from the nucleus to the cytoplasm and for translation of the mRNA.

Poly-A tails are also found on mRNAs in bacteria and archaea. However, these bacterial poly-A tails are generally much shorter and found on only a small fraction of mRNA molecules. In addition, whereas poly-A tails are protective in eukaryotes, poly-A tails are generally associated with mRNA degradation in bacteria.

13.11 The Coding Regions of Eukaryotic Genes Are Interrupted by Intervening Sequences Called Introns

As mentioned above, the primary mRNA transcript, or pre-mRNA, is often longer than the mature mRNA in eukaryotes. An explanation for this phenomenon emerged in 1977 when research groups led by Phillip Sharp and Richard Roberts independently published direct evidence that the genes of animal viruses contain *internal* (also referred to as *intervening* or *intragenic*) nucleotide sequences that do not encode for amino acids in the final protein product. These noncoding internal sequences are also present in pre-mRNAs, but they are removed during RNA processing to produce the mature mRNA (Figure 13.10), which is then translated. Such nucleotide sequences—ones that intervene between sequences that code for amino acids—are called **introns** (derived from *intr*agenic regi*on*). Sequences that are retained in the mature mRNA and expressed are called **exons** (for *ex*pressed regi*on*). The process of removing introns from a pre-mRNA and joining together exons is called **RNA splicing**.

One of the first intron-containing genes identified was the **β-globin gene** in mice and rabbits, studied independently by Philip Leder and Richard Flavell. The mouse gene contains an intron 550 nucleotides long, beginning immediately after the sequence specifying the 104th amino acid. In rabbits, there is an intron of 580 base pairs near the sequence for the 110th amino acid—a strikingly similar pattern to that seen in mice. In addition, another intron of about 120 nucleotides exists earlier in both genes. Similar introns have been found in the β-globin gene in all mammals examined thus far.

The **ovalbumin gene** of chickens has been extensively characterized by Bert O'Malley in the United States and Pierre Chambon in France. As shown in **Figure 13.12**, the gene contains seven introns. In fact, the majority of the gene's DNA sequence is composed of introns and is thus noncoding. The pre-mRNA is nearly three times the length of the mature spliced mRNA.

The identification of introns in eukaryotic genes involves a direct comparison of nucleotide sequences of DNA with those of mRNA and their correlation with amino acid sequences. Such an approach allows the precise identification of all intervening sequences. By identifying common sequences that appear at intron/exon boundaries, scientists are now able to identify introns with excellent accuracy using only the genomic DNA sequence and computational tools. We will return to this topic when we consider genomic analysis (Chapter 21).

We now have a fairly comprehensive view of intron-containing eukaryotic genes from many species. In the budding yeast *Saccharomyces cervisiae*, 283 out of the roughly 6000 protein-coding genes have introns. However, introns are far more common in humans; roughly 94 percent of human protein-coding genes contain introns with an average of nine exons and eight introns per gene. An extreme example of the number of introns present in a single gene is provided by the gene coding for one of the subunits of collagen, the major connective tissue protein in vertebrates. The *pro-α-2(1) collagen* gene contains 51 introns. The precision of RNA splicing must be extraordinary if errors are not to be introduced into the mature mRNA.

Equally noteworthy is the difference between the length of a typical gene and the length of the final mRNA after introns are removed by splicing. As shown in

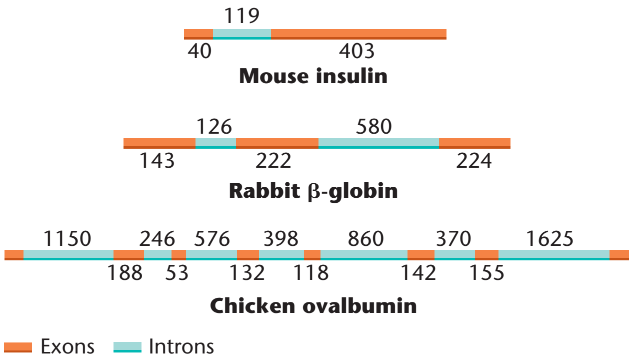


FIGURE 13.12 Intron and exon sequences in various eukaryotic genes. The numbers indicate the number of nucleotides present in various intron and exon regions.

TABLE 13.7 Contrasting Human Gene Size, mRNA Size, and Number of Introns

Gene	Gene Size (kb)	mRNA Size (kb)	Number of Introns
Insulin	1.7	0.4	2
Collagen [*pro-α-2(1)*]	38.0	5.0	51
Albumin	25.0	2.1	14
Phenylalanine hydroxylase	90.0	2.4	12
Dystrophin	2400.0	17.0	79

Table 13.7, only about 13 percent of the collagen gene consists of exons that appear in mature mRNA. For other genes, an even more extreme picture emerges. Only about 8 percent of the albumin gene codes for the amino acids in the albumin protein, and in the largest human gene known, dystrophin (which is the protein product absent in Duchenne muscular dystrophy), less than 1 percent of the gene sequence is retained in the mRNA.

Although the vast majority of mammalian genes examined thus far contain introns, there are several exceptions. Notably, the genes coding for histones and for interferon, a signaling protein of the immune system, appear to contain no introns.

Why Do Introns Exist?

A curious genetics student who first learns about the concept of introns and RNA splicing often wonders why introns exist. If intron sequences are destined for removal, then why are they there in the first place? Wouldn't it be more efficient if introns were absent and hence never transcribed? Indeed, scientists asked these same questions shortly after introns were discovered in 1977. However, we know now that introns indeed serve several functions:

1. Some genes can encode for more than one protein product through the alternative use of exons. This process, known as **alternative splicing** (described in more detail in Chapter 18), produces different mature mRNAs from the same pre-mRNA by splicing out introns and ligating together different combinations of exons. This means that a eukaryotic genome can encode a greater number of proteins than it has protein-coding genes.

2. Introns may also be important to the evolution of genes. On evolutionary time scales, DNA sequences may be moved around within the genome. The modular exon/intron gene structure allows for a phenomenon known as **exon shuffling** (described in more detail in Chapter 14), whereby new genes may evolve when an exon is introduced into an existing gene.

3. Once an intron is excised from a pre-mRNA, it is generally degraded. However, there are many documented cases where an intron actually contains a noncoding RNA, such as a **microRNA (miRNA)**, which is a small RNA that regulates gene expression (see Chapter 18). In such cases, the excised intron is processed to liberate the noncoding RNA, which then functions within the cell.

4. Introns can also regulate transcription. For example, intronic sequences in the DNA frequently harbor *cis* regulatory elements, such as enhancers and silencers that upregulate or downregulate transcription, respectively.

Splicing Mechanisms: Self-Splicing RNAs

The discovery of introns led to intensive attempts to elucidate the mechanism by which they are excised and exons are spliced back together. A great deal of progress has been made, relying heavily on *in vitro* studies. Interestingly, it appears that somewhat different mechanisms exist for different classes of transcripts, as well as for RNAs produced in mitochondria and chloroplasts.

We might envision the simplest possible mechanism for removing an intron to involve two steps: (1) the intron is cut at both ends by an endonuclease and (2) the adjacent exons are joined, or ligated, by a ligase. This is, apparently, what happens to the introns present in transfer RNAs (tRNAs) in bacteria. However, in studies of other RNAs—tRNAs in higher eukaryotes and rRNAs and pre-mRNAs in all eukaryotes—precise excision of introns is much more complex and a much more interesting story.

Introns in eukaryotes can be categorized into several groups based on their splicing mechanisms. *Group I introns*, such as those in the primary transcript of rRNAs, require no outside help for intron excision; the intron itself is the source of the enzymatic activity necessary for splicing. This amazing discovery was made in 1982 by Thomas Cech and colleagues during a study of the ciliate protozoan *Tetrahymena*. RNAs that are capable of such catalytic activity are referred to as **ribozymes**.

The self-excision process for group I introns is illustrated in **Figure 13.13**. Chemically, two nucleophilic reactions take place—that is, reactions caused by the presence of electron-rich chemical species (in this case, they are *transesterification reactions*). The first is an interaction between a free guanosine (symbolized as "G"), which acts as a cofactor in the reaction, and the primary transcript [**Figure 13.13(a)**]. After guanosine is positioned in the active site of the intron, its 3′-OH group attacks and breaks the phosphodiester bond ("P") between the nucleotides at the 5′ end of the intron and the 3′ end of the left-hand exon [**Figure 13.13(b)**]. The second reaction involves the interaction of the newly formed 3′-OH group on the left-hand exon and the phosphodiester bond at the right intron/exon boundary [**Figure 13.13(c)**]. The intron is spliced out and the two exons are ligated, leading to the mature RNA [**Figure 13.13(d)**].

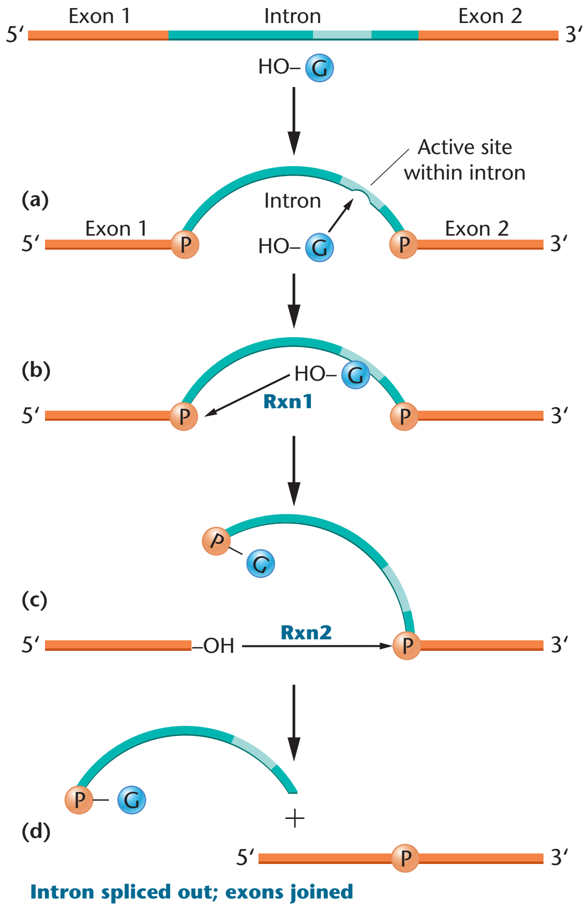


FIGURE 13.13 Splicing mechanism for removal of a group I intron. The process is one of self-excision involving two transesterification reactions.

Self-excision of group I introns, as described above, is known to occur in preliminary transcripts for mRNAs, tRNAs, and rRNAs in bacteria, lower eukaryotes, and higher plants.

Self-excision also governs the removal of introns from the primary mRNA and tRNA transcripts produced in mitochondria and chloroplasts; these are examples of *group II introns*. Splicing of group II introns is somewhat different than for group I, but also involves two autocatalytic reactions leading to the excision of introns. Group II introns are found in fungi, plants, protists, and bacteria.

Splicing Mechanisms: The Spliceosome

Compared to the group I and group II introns discussed above, *those in nuclear-derived protein-coding pre-mRNAs* are much larger—several examples exceed 500,000 nucleotides—and they are more plentiful. Their removal appears to require a much more complex mechanism. These splicing reactions are not autocatalytic, but instead are mediated by a molecular complex called the **spliceosome.** This structure is very large, 40*S* in yeast and 60*S* in mammals, being the same size as ribosomal subunits! Introns removed by the spliceosome are known as *spliceosomal introns*.

One set of essential components of spliceosomes are the **small nuclear RNAs (snRNAs).** These RNAs are usually 80 to 400 nucleotides long and, because they are rich in uridine residues, have been arbitrarily named U1, U2, . . . , U6. The snRNAs are complexed with proteins to form **small nuclear ribonucleoproteins (snRNPs),** pronounced "snurps," which are named after the specific snRNAs contained within them (the U2 snRNA is contained within the U2 snRNP).

Figure 13.14 depicts a model of the steps involved in the removal of one spliceosomal intron. Keep in mind that

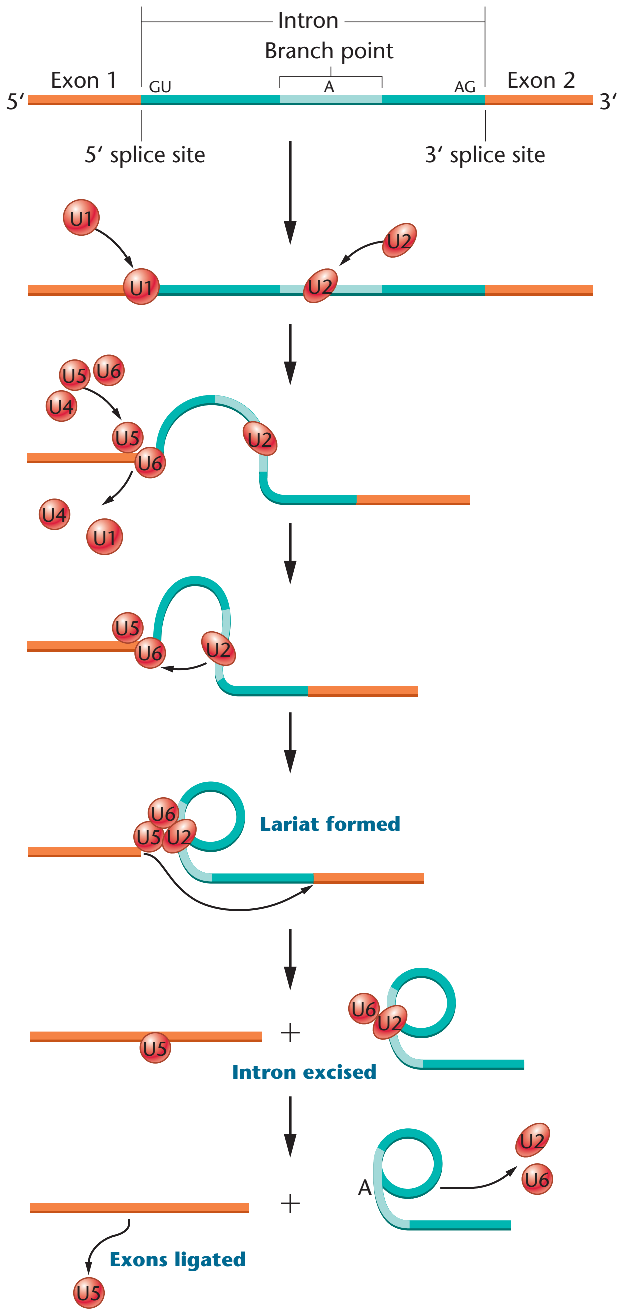


FIGURE 13.14 A model of the splicing mechanism for removal of a spliceosomal intron. Excision is dependent on snRNPs (U1, U2, etc.). The lariat structure is characteristic of this mechanism.

while this figure represents snRNPs as separate components, they are actually part of the huge spliceosome that envelops the RNA being spliced. The nucleotide sequences near the ends of the intron begin at the 5′ end with a GU dinucleotide sequence, called the **splice donor** sequence, and terminate at the 3′ end with an AG dinucleotide, called the **splice acceptor** sequence. These, as well as other consensus sequences shared by spliceosomal introns, attract specific snRNAs of the spliceosome. For example, the U1 snRNA bears a nucleotide sequence that is complementary to the 5′-splice donor sequence. Base pairing resulting from this complementarity promotes the binding that represents the initial step in the formation of the spliceosome. After the other snRNPs (U2, U4, U5, and U6) are added, splicing commences.

As with group I splicing, two *transesterification reactions* occur. The first involves an adenine (A) residue present within the **branch point** region of the intron. The 2′-OH of this A residue attacks the phosphodiester bond at the 5′-splice site, cutting the RNA chain and forming an atypical 5′-to-2′ bond between the 5′ end of the intron and the branch point A residue. In the next step, the newly formed 3′-OH of the upstream exon attacks the phosphodiester bond at the 3′-splice site thus excising the intron and joining the two exons. The excised intron has a characteristic loop structure, called a *lariat*, due to the 5′-to-2′ bond produced in the first transesterification reaction. Once the second reaction is complete, the snRNPs are released and the excised intron is generally degraded by nucleases.

Recently, a study by the Staley and Piccirilli groups at the University of Chicago demonstrated that specific phosphates within the U6 snRNA bind and position two magnesium ions required for both catalytic steps shown in Figure 13.14. This strongly suggests that the spliceosome is in fact a ribozyme—a catalytic RNA. Furthermore, when the 3D structure of the U2/U6 snRNP complex is compared to the 3D structure of group II self-splicing introns, the specific metal-binding phosphates in U6 correspond to known metal-binding phosphates in group II introns. This substantiates a long-held belief that group II self-splicing introns are the evolutionary ancestor of the modern spliceosome. RNA splicing within the nucleus, as described above, represents a potential regulatory step in gene expression in eukaryotes. For instance, many cases are known wherein introns present in nuclear pre-mRNAs *derived from the same gene* are spliced *in more than one way*, thereby yielding different collections of exons in the mature mRNAs. Such *alternative splicing* yields a group of similar but nonidentical mRNAs that, upon translation, result in a set of related proteins called **isoforms.** Many examples have been encountered in organisms including plants, *Drosophila,* and humans. Alternative splicing represents a way of producing related proteins from a single gene, increasing the number of gene products that can be derived from an organism's genome. We will return to this topic in our discussion of the posttranscriptional regulation of gene expression in eukaryotes (see Chapter 18).

EVOLVING CONCEPT OF THE GENE

The elucidation of the genetic code in the 1960s supported the concept that the gene is composed of an uninterrupted linear series of triplet nucleotides encoding the amino acid sequence of a protein. While this is indeed the case in bacteria and viruses that infect them, in 1977, it became apparent that in eukaryotes, the gene is divided into coding sequences, called exons, which are interrupted by noncoding sequences, called introns, which must be spliced out during production of the mature mRNA. ■

13.12 RNA Editing May Modify the Final Transcript

In the late 1980s, still another unexpected form of posttranscriptional RNA processing was discovered in several organisms. In this form, referred to as **RNA editing,** the ribonucleotide sequence of a pre-mRNA is actually changed prior to translation. As a result, the ribonucleotide sequence of the mature RNA differs from the sequence encoded in the exons of the DNA from which the RNA was transcribed.

Although other variations exist, there are two main types of RNA editing: **substitution editing,** in which the identities of individual nucleotide bases are altered via chemical modification; and **insertion/deletion editing,** in which nucleotides are added to or subtracted from the total number of bases. Substitution editing is used in some nuclear-derived eukaryotic RNAs and is prevalent in mitochondrial and chloroplast RNAs transcribed in plants. *Physarum polycephalum*, a slime mold, uses both substitution and insertion/deletion editing for its mitochondrial mRNAs.

Trypanosoma, a genus of parasite that causes African sleeping sickness, and its relatives use extensive insertion/deletion editing in mitochondrial RNAs. Uridines added to an individual transcript can make up more than 60 percent of the coding sequence, usually forming the initiation codon and bringing the rest of the sequence into the proper reading frame. Insertion/deletion editing in trypanosomes is directed by **gRNA (guide RNA)** templates, which are also transcribed from the mitochondrial genome. These small RNAs share a high degree of complementarity to the edited region of the final mRNAs. They base-pair with the pre-edited mRNAs to direct the editing machinery to make the correct changes.

The best-studied examples of substitutional editing occur in mammalian nuclear-encoded mRNA transcripts. One such example is the protein *apolipoprotein B* (*ApoB*), which exists in both a long and a short form that are encoded by the *APOB* gene. In human intestinal cells, *APOB* mRNA editing produces a single C-to-U change, which converts a CAA glutamine codon into a UAA stop codon and terminates the polypeptide at approximately half its genomically encoded length. In this example, RNA editing is performed by a complex of proteins, including Apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1 (APOBEC-1), which bind to a "mooring sequence" on the pre-mRNA transcript just downstream of the editing site. **Figure 13.15(a)** shows the reaction catalyzed during this editing event, in which deamination of cytidine yields uridine.

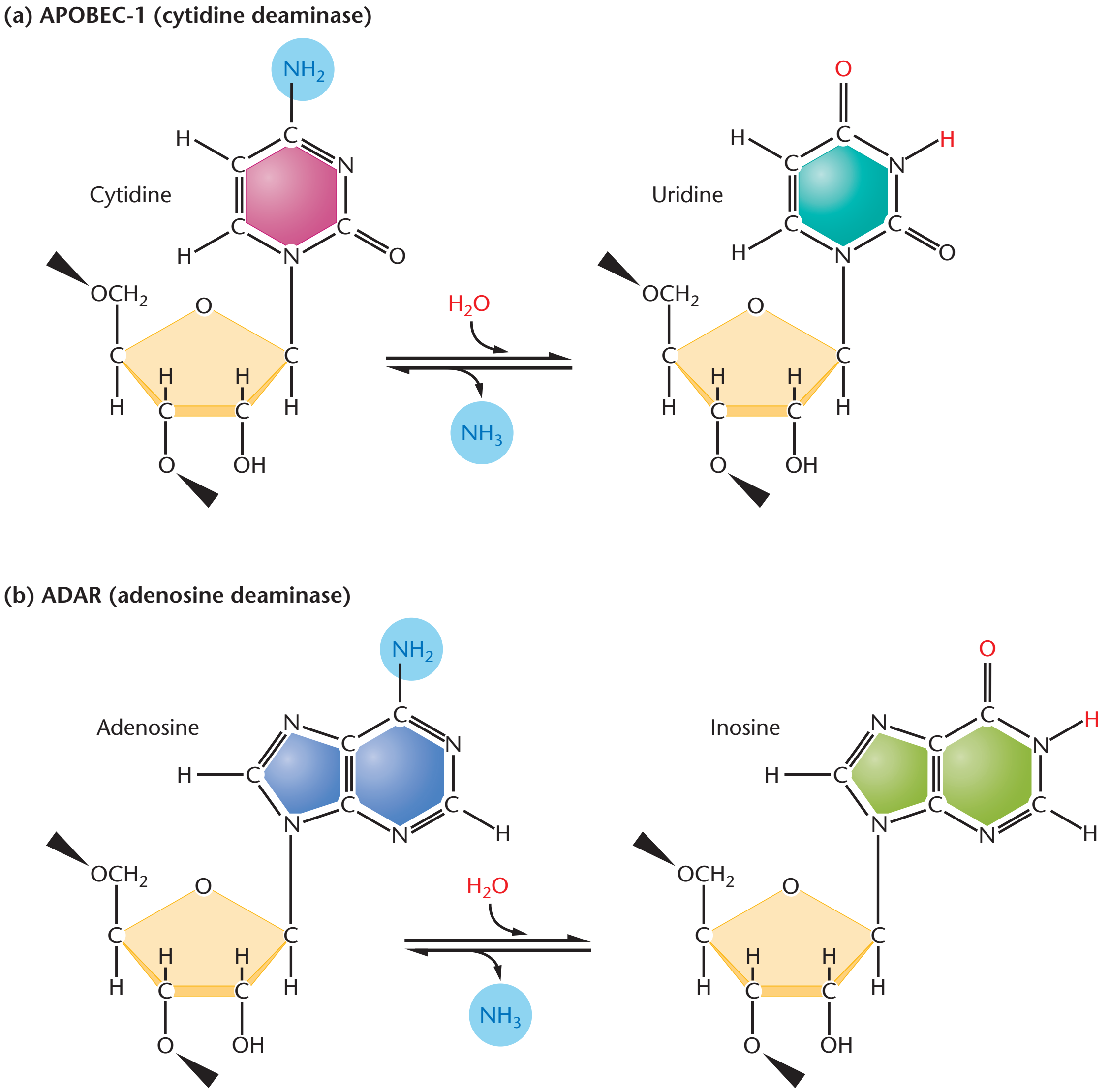

(a) APOBEC-1 (cytidine deaminase)
Cytidine
Uridine
NH2
H2O
NH3
OCH2
O
C
N
H
OH
(b) ADAR (adenosine deaminase)
Adenosine
Inosine
NH2
H2O
NH3
OCH2

FIGURE 13.15 RNA editing reactions: Deamination of cytidine by the enzyme APOBEC-1 results in uridine (a), whereas deamination of adenosine by the enzyme ADAR produces the noncanonical nucleoetide, inosine (b).

A second example involves the subunits constituting the *glutamate receptor channels (GluR)* in mammalian brain tissue. In this case, adenosine (A) to inosine (I) editing occurs in pre-mRNAs. During translation, the substituted I is read as guanosine (G). A family of three ADAR (*a*denosine *d*eaminase *a*cting on *R*NA) enzymes is believed to be responsible for the editing of various sites within the glutamate channel subunits. **Figure 13.15(b)** depicts ADAR-mediated deamination of adenosine to produce inosine.

Double-stranded RNA is required for editing by the ADAR enzymes. Double-stranded regions are provided by intron/exon base pairing within the GluR pre-mRNA transcripts. Expression of edited and unedited GluR subunits in cultured cells showed that the changes catalyzed by ADAR ultimately alter the physiological parameters (solute permeability and desensitization response time) of the receptors containing the subunits encoded by the edited RNAs.

A similar ADAR-mediated editing step has more recently been reported in a voltage-dependent potassium (K^+) channel in a cephalopod. The channel regulates neuron function. In this interesting example, editing of channel-encoding mRNA occurs in squid living in near-freezing water, but not in the same species found living in more tropical environments. Thus, RNA editing has been shown to be adaptive, in this case responding to cold temperatures.

Findings such as these have established that RNA editing provides still another important mechanism of posttranscriptional modification. These discoveries, too, have important implications for the regulation of genetic expression.

13.13 Transcription Has Been Visualized by Electron Microscopy

We conclude this chapter by presenting a striking visual demonstration of the transcription process based on classic electron microscope studies by Oscar Miller, Jr., Barbara Hamkalo, and Charles Thomas. Their combined work has captured the transcription process in both bacteria and eukaryotes. **Figure 13.16** shows a micrograph and interpretive drawings of transcription in *E. coli*. In the micrograph, multiple strands of RNA are seen to emanate from different points along a DNA template. Many RNA strands result because numerous transcription events are occurring simultaneously. Progressively longer RNA strands are found farther downstream from the point of initiation of transcription along a given gene, whereas the shortest strands are closest to the point of initiation.

An interesting picture emerges from study of the *E. coli* micrograph. Because bacteria lack nuclei, cytoplasmic ribosomes are not separated physically from the chromosome. As a result, ribosomes are free to attach to *partially* transcribed mRNA molecules and initiate translation. The longer the RNA strand, the greater the number of ribosomes attached to it—as we will see later in the text (Chapter 14). These structures are called **polyribosomes.** Visualization of transcription confirms many of the predictions scientists had made from biochemical analysis of this process.

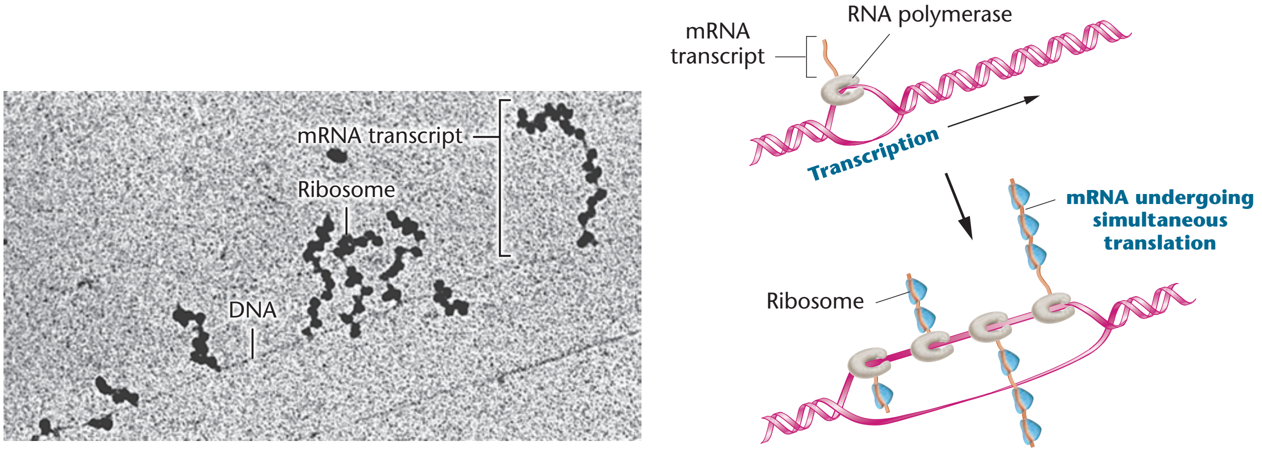


FIGURE 13.16 Electron micrograph and interpretive drawings of simultaneous transcription and translation of genes in *E. coli*. As each mRNA transcript is forming, ribosomes attach, initiating translation along each strand.

CASE STUDY Treatment dilemmas

A 30-year- old woman was undergoing therapy for β-thalassemia, a recessive trait caused by absence of or reduced synthesis of the hemoglobin β chain, a subunit of the oxygen-carrying molecule in red blood cells. In this condition, red blood cells are rapidly destroyed, freeing a large amount of iron, which is deposited in tissues and organs. The blood transfusions the patient had received every two or three weeks since the age of 7 to stave off anemia were further aggravating iron buildup. Her major organs were showing damage, and she was in danger of death from cardiac disease. Her physician suggested that she consider undergoing a hematopoietic (bone marrow) stem cell transplant (HSCT). Since these stem cells give rise to red blood cells, such a transplant could potentially restore her health. While this might seem like an easy decision, it is not. Advanced cases have a high risk (almost 30 percent) for transplantation-related death. At this point, the woman is faced with a difficult and important decision.

1. Consider different ways in which a mutation, a single base-pair change or small deletion, in the gene encoding hemoglobin β chain could lead to β-thalassemia. For example, how might mutations in promoter, enhancer, or coding regions yield this outcome?
2. Why is it important that the physician emphasize to the patient that she must bear the responsibility for the final decision (i.e., that once she has considered all aspects of the decision, she act autonomously)?
3. If you were faced with this decision, what further input might you seek?

For related reading, see Caocci, G., et al. (2011). Ethical issues of unrelated hematopoietic stem cell transplantation in adult thalassemia patients. *BMC Medical Ethics* 12(4):1–7.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Early studies of the genetic code revealed it to be triplet in nature and to be nonoverlapping, commaless, and degenerate.
2. The use of RNA homopolymers and heteropolymers in a cell-free protein-synthesizing system allowed the determination of the composition, but not the sequence, of triplet codons designating specific amino acids.
3. Use of the triplet binding assay and of repeating copolymers allowed the determination of the sequences of triplet codons designating specific amino acids.
4. The complete coding dictionary reveals that of the 64 possible triplet codons, 61 encode the 20 amino acids found in proteins, while three triplets terminate translation. One of the 61 amino-acid–coding triplets is the initiation codon and specifies methionine.
5. Confirmation for the coding dictionary, including codons for initiation and termination, was obtained by comparing the complete nucleotide sequences of phage MS2 with the amino acid sequence of the corresponding proteins. Other findings support the belief that, with only minor exceptions, the code is universal for all organisms.
6. Transcription—the initial step in gene expression—is the synthesis, under the direction of RNA polymerase, of a strand of RNA complementary to a DNA template.
7. Like DNA replication, the processes of transcription can be subdivided into the stages of initiation, elongation, and termination. Also like DNA replication, transcription relies on base-pairing affinities between complementary nucleotides.
8. Initiation of gene transcription is dependent on an upstream (5′) DNA region, called a promoter, which represents the initial binding site for RNA polymerase. Promoters contain specific DNA sequences, such as the -10 site (Pribnow box) in bacteria and the TATA box in eukaryotes, which are essential to polymerase binding.
9. Whereas bacterial mRNAs are able to be translated immediately, the initial transcripts of eukaryotes must undergo RNA processing steps that include adding a 5′ m^7G cap, splicing out noncoding sequences called introns, and adding a 3′ poly-A tail to make the mature mRNA.
10. Some introns are self-splicing and catalyze their own removal from the primary transcript. However, most eukaryotic introns are removed by an RNA/protein complex called the spliceosome.
11. RNA editing alters the ribonucleotide sequence of an mRNA molecule prior to its translation.

GENETICS, ETHICS, AND SOCIETY

Treating Duchenne Muscular Dystrophy with Exon-Skipping Drugs

One in every 3500 newborn males is afflicted by a serious X-linked recessive disease known as Duchenne muscular dystrophy (DMD). This disease is progressive, resulting in muscle degeneration, heart disease, and premature death.

DMD is caused by mutations in the *dystrophin* gene, a 2400-kb gene containing 79 exons. The DMD primary transcript is 2100 kb long and takes about 16 hours to transcribe. Many DMD mutations are frameshift mutations that shift the reading frame in the mRNA so that at least one codon downstream of the frameshift mutation becomes a stop codon. This stop codon causes

(continued)

Genetics, Ethics, and Society, continued

premature termination of translation of the mRNA. Most of the resulting truncated dystrophin proteins are not functional. There are no cures for DMD and few effective treatments.

Recently, a new DMD drug called eteplirsen completed clinical trials and received accelerated approval by the U.S. Food and Drug Administration (FDA). This drug uses a technique called *exon skipping* to target mutations in exon 51 of *dystrophin*.

In exon skipping, an exon that bears a mutation is targeted and removed during pre-mRNA splicing. If the exons that precede and follow the mutated exon are spliced together in such a way that the reading frame is restored, the resulting protein may retain some activity, even though it lacks the amino acids encoded by the skipped exon.

Eteplirsen is a molecule known as an *antisense oligonucleotide* (*ASO*). ASOs are short synthetic single-stranded DNA molecules that have specific sequences complementary to a portion of a targeted mRNA. As mRNA molecules are transcribed from the template strand of DNA, they are known as *sense* RNAs. The ASOs, being complementary to mRNA sequences, are therefore *antisense* molecules. When ASOs enter cells, they bind to their complementary sequence in the target mRNA. This results in the mRNA's degradation, or interference with its splicing or translation.

Once inside a cell, eteplirsen binds to the *dystrophin* pre-mRNA exon 51 splice junctions, interfering with normal pre-mRNA splicing. As a result, exon 50 is spliced to exon 52, thereby eliminating exon 51 and its mutation from the mature mRNA. This internal deletion restores the correct reading frame to the *dystrophin* mRNA. Upon translation, the new dystrophin protein, although lacking some amino acids, has enough activity to restore partial function to the patient's muscles.

Treatment of DMD with eteplirsen has been so encouraging that other exon-skipping drugs are being developed for DMD patients who have mutations in other exons, including exons 8, 44, 50, and 55.

Your Turn

Take time, individually or in groups, to consider the ethical and technical issues that surround the development and uses of DMD exon-skipping drugs.

1. In September 2016, the U.S. FDA gave accelerated approval to eteplirsen. Despite the fact that the FDA's advisory panel voted against approval, intense lobbying by DMD patients and their families may have contributed to the speed at which the FDA approved this new drug. Discuss this case of accelerated approval, considering ethical arguments on both sides of the controversy. Do you think that the FDA's approval of eteplirsen was justified?

   Read about the FDA's approval of eteplirsen at: Tavernise, S., F.D.A. approves muscular dystrophy drug that patients lobbied for. *New York Times*, September 19, 2016 (http://www.nytimes.com/2016/09/20/business/fda-approves-muscular-dystrophy-drug-that-patients-lobbied-for.html).

2. Only about 13 percent of DMD patients have exon 51 mutations leading to premature translation termination. To target other mutated exons, several biotechnology companies are researching ASOs that could cause exon skipping in other parts of the *dystrophin* mRNA. Describe these new drugs and their modes of action. Are any of these new drugs in clinical trials?

   Investigate the links on the Muscular Dystrophy Association Web site (https://www.mda.org/quest/article/exon-skipping-dmd-what-it-and-whom-can-it-help).

INSIGHTS AND SOLUTIONS

1. Calculate how many triplet codons would be possible had evolution seized on six bases (three complementary base pairs) rather than four bases with which to construct DNA. Would six bases accommodate a two-letter code, assuming 20 amino acids and start and stop codons?

   Solution: Six bases taken three at a time would produce $(6)^3$, or 216, triplet codes. If the code was a doublet, there would be $(6)^2$, or 36, two-letter codes, more than enough to accommodate 20 amino acids and start and stop signals.

2. In a heteropolymer experiment using 1/2C : 1/4A : 1/4G, how many different triplets will occur in the synthetic RNA molecule? How often will the most frequent triplet occur?

   Solution: There will be $(3)^3$, or 27, triplets produced. The most frequent will be CCC, present $(1/2)^3$ or 1/8 of the time.

3. In a copolymer experiment in which the tetranucleotide sequence UUAC is repeated over and over, how many different triplets will occur in the synthetic RNA, and how many amino acids will occur in the polypeptide when this RNA is translated? (Consult Figure 13.7.)

   Solution: The synthetic RNA will repeat four triplets—UUA, CUU, ACU, and UAC—over and over. Because both UUA and CUU encode leucine, while ACU and UAC encode threonine and tyrosine, respectively, the polypeptides synthesized under the directions of such an RNA contain three amino acids in the repeating sequence Leu-Leu-Thr-Tyr.

4. Actinomycin D inhibits DNA-dependent RNA synthesis. This antibiotic is added to a bacterial culture in which a specific protein is being monitored. Compared to a control culture, into which no antibiotic is added, translation of the protein declines over a period of 20 minutes, until no further protein is made. Explain these results.

   Solution: The mRNA, which is the basis for the translation of the protein, has a lifetime of about 20 minutes. When actinomycin D is added, transcription is inhibited and no new mRNAs are made. Those already present support the translation of the protein for up to 20 minutes.

Problems and Discussion Questions

1. **HOW DO WE KNOW?** In this chapter, we focused on the genetic code and the transcription of genetic information stored in DNA into complementary RNA molecules. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
(a) Why did geneticists believe, even before direct experimental evidence was obtained, that the genetic code would turn out to be composed of triplet sequences and be nonoverlapping? Experimentally, how were these suppositions shown to be correct?
(b) What experimental evidence provided the initial insights into the *compositions* of codons encoding specific amino acids?
(c) How were the specific sequences of triplet codes determined experimentally?
(d) How were the experimentally derived triplet codon assignments verified in studies using bacteriophage MS2?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p. 283. These all center around how genetic information is stored in DNA and transferred to RNA prior to translation into proteins. Write a short essay that summarizes the key properties of the genetic code and the process by which RNA is transcribed on a DNA template.

3. Assuming the genetic code is a triplet, what effect would the addition or loss of two nucleotides have on the reading frame? The addition or loss of three, six, or nine nucleotides?

4. The mRNA formed from the repeating tetranucleotide UUAC incorporates only three amino acids, but the use of UAUC incorporates four amino acids. Why?

5. In studies using repeating copolymers, AC . . . incorporates threonine and histidine, and CAACAA . . . incorporates glutamine, asparagine, and threonine. What triplet code can definitely be assigned to threonine?

6. In a coding experiment using repeating copolymers (as demonstrated in Table 13.3), the following data were obtained:

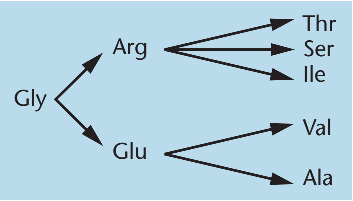
Copolymer	Codons Produced	Amino Acids in Polypeptide
AG	AGA, GAG	Arg, Glu
AAG	AGA, AAG, GAA	Lys, Arg, Glu

AGG is known to code for arginine. Taking into account the wobble hypothesis, assign each of the four codons produced in the experiment to its correct amino acid.

7. In the triplet binding technique, radioactivity remains on the filter when the amino acid corresponding to the codon is labeled. Explain the rationale for this technique.

8. When the amino acid sequences of insulin isolated from different organisms were determined, differences were noted. For example, alanine was substituted for threonine, serine for glycine, and valine for isoleucine at corresponding positions in the protein. List the single-base changes that could occur in codons of the genetic code to produce these amino acid changes.

9. In studies of the amino acid sequence of wild-type and mutant forms of tryptophan synthetase in *E. coli*, the following changes have been observed:



Determine a set of triplet codes in which only a single-nucleotide change produces each amino acid change.

10. Why doesn't polynucleotide phosphorylase (Ochoa's enzyme) synthesize RNA *in vivo*?

11. Refer to Table 13.1. Can you hypothesize why a synthetic RNA composed of a mixture of poly U + poly A would not stimulate incorporation of ^{14}C-phenylalanine into protein?

12. Predict the amino acid sequence produced during translation by the following short hypothetical mRNA sequences (note that the second sequence was formed from the first by a deletion of only one nucleotide):

Sequence 1: 5′-AUGCCGGAUUAUAGUUGA-3′
Sequence 2: 5′-AUGCCGGAUUAAGUUGA-3′

What type of mutation gave rise to sequence 2?

13. A short RNA molecule was isolated that demonstrated a hyperchromic shift (see Chapter 10), indicating secondary structure. Its sequence was determined to be

5′-AGGCGCCGACUCUACU-3′

(a) Propose a two-dimensional model for this molecule.
(b) What DNA sequence would give rise to this RNA molecule through transcription?
(c) If the molecule were a tRNA fragment containing a CGA anticodon, what would the corresponding codon be?
(d) If the molecule were an internal part of a message, what amino acid sequence would result from it following translation? (Refer to the code chart in Figure 13.7.)

14. A glycine residue is in position 210 of the tryptophan synthetase enzyme of wild-type *E. coli*. If the codon specifying glycine is GGA, how many single-base substitutions will result in an amino acid substitution at position 210? What are they? How many will result if the wild-type codon is GGU?

15. Refer to Figure 13.7 to respond to the following:
(a) Shown here is a hypothetical viral mRNA sequence:

5′-AUGCAUACCUAUGAGACCCUUGGA-3′

Assuming that it could arise from overlapping genes, how many different polypeptide sequences can be produced? What are the sequences?
(b) A base-substitution mutation that altered the sequence shown in part (a) eliminated the synthesis of all but one polypeptide. The altered sequence is shown here:

5′-AUGCAUACCUAUGUGACCCUUGGA-3′

Determine why.

16. Most proteins have more leucine than histidine residues, but more histidine than tryptophan residues. Correlate the number of codons for these three amino acids with this information.

17. Define the process of transcription. Where does this process fit into the central dogma of molecular biology (DNA makes RNA makes protein)?

18. What observations suggested the existence of mRNA?

19. Describe the structure of RNA polymerase in bacteria. What is the core enzyme? What is the role of the σ subunit?

20. Write a paragraph describing the abbreviated chemical reactions that summarize RNA polymerase-directed transcription.

21. Messenger RNA molecules are very difficult to isolate in bacteria because they are rather quickly degraded in the cell. Can you suggest a reason why this occurs? Eukaryotic mRNAs are more stable and exist longer in the cell than do bacterial mRNAs. Is this an advantage or a disadvantage for a pancreatic cell making large quantities of insulin?

22. Present an overview of various forms of posttranscriptional RNA processing in eukaryotes. For each, provide an example.

23. One form of posttranscriptional modification of most eukaryotic pre-mRNAs is the addition of a poly-A sequence at the 3′ end. The absence of a poly-A sequence leads to rapid degradation of the transcript. Poly-A sequences of various lengths are also added to many bacterial RNA transcripts where, instead of promoting stability, they enhance degradation. In both cases, RNA secondary structures, stabilizing proteins, or degrading enzymes interact with poly-A sequences. Considering the activities of RNAs, what might be general functions of 3′-polyadenylation?

24. Describe the role of two forms of RNA editing that lead to changes in the size and sequence of pre-mRNAs. Briefly describe several examples of each form of editing, including their impact on respective protein products.

25. Substitution RNA editing is known to involve either C-to-U or A-to-I conversions. What common chemical event accounts for each?

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

26. It has been suggested that the present-day triplet genetic code evolved from a doublet code when there were fewer amino acids available for primitive protein synthesis.
(a) Can you find any support for the doublet code notion in the existing coding dictionary?
(b) The amino acids Ala, Val, Gly, Asp, and Glu are all early members of biosynthetic pathways and are more evolutionarily conserved than other amino acids. They therefore probably represent "early" amino acids. Of what significance is this information in terms of the evolution of the genetic code? Also, which base, of the first two within a coding triplet, would likely have been the more significant in originally specifying these amino acids?
(c) As determined by comparisons of ancient and recently evolved proteins, cysteine, tyrosine, and phenylalanine appear to be late-arriving amino acids. In addition, they are considered to have been absent in the abiotic Earth. All three of these amino acids have only two codons each, while many others, earlier in origin, have more. Is this mere coincidence, or might there be some underlying explanation?

27. An early proposal by George Gamow in 1954 regarding the genetic code considered the possibility that DNA served directly as the template for polypeptide synthesis. In eukaryotes, what difficulties would such a system pose? What observations and theoretical considerations argue against such a proposal?

28. In a mixed copolymer experiment, messages were created with either 4/5C : 1/5A or 4/5A : 1/5C. These messages yielded proteins with the following amino acid compositions.

	4/5C : 1/5A (% yield)	4/5A : 1/5C (% yield)
Proline	63.0	3.5
Histidine	13.0	3.0
Threonine	16.0	16.6
Glutamine	3.0	13.0
Asparagine	3.0	13.0
Lysine	0.5	50.0
	98.5	99.1

Using these data, predict the most specific coding composition for each amino acid.

29. Shown here are the amino acid sequences of the wild-type and three mutant forms of a short protein.

Wild-type:	Met-Trp-Tyr-Arg-Gly-Ser-Pro-Thr
Mutant 1:	Met-Trp
Mutant 2:	Met-Trp-His-Arg-Gly-Ser-Pro-Thr
Mutant 3:	Met-Cys-Ile-Val-Val-Val-Gln-His

Use this information to answer the following questions:
(a) Using Figure 13.7, predict the type of mutation that led to each altered protein.
(b) For each mutant protein, determine the specific ribonucleotide change that led to its synthesis.
(c) The wild-type RNA consists of nine triplets. What is the role of the ninth triplet?
(d) Of the first eight wild-type triplets, which, if any, can you determine specifically from an analysis of the mutant proteins? In each case, explain why or why not.
(e) Another mutation (mutant 4) is isolated. Its amino acid sequence is unchanged from wild type, but the mutant cells produce abnormally low amounts of the wild-type proteins. As specifically as you can, predict where this mutation exists in the gene.

30. The genetic code is degenerate. Amino acids are encoded by either 1, 2, 3, 4, or 6 triplet codons (see Figure 13.7). An interesting question is whether the number of triplet codes for a given amino acid is in any way correlated with the frequency with which that amino acid appears in proteins. That is, is the genetic code optimized for its intended use? Some approximations of the frequency of appearance of nine amino acids in proteins in *E. coli* are given in the following:

Amino Acid	Percentage
Met	2
Cys	2
Gln	5
Pro	5
Arg	5
Ile	6
Glu	7
Ala	8
Leu	10

(a) Determine how many triplets encode each amino acid.
(b) Devise a way to graphically compare the two sets of information (data).
(c) Analyze your data to determine what, if any, correlations can be drawn between the relative frequency of amino acids making up proteins and the number of codons for each. Write a paragraph that states your specific and general conclusions.
(d) How would you proceed with your analysis if you wanted to pursue this problem further?

31. M. Klemke et al. (2001) discovered an interesting coding phenomenon in which an exon within a neurologic hormone receptor gene in mammals appears to produce two different protein entities (XLαs and ALEX). Following is the DNA sequence of the exon's 5′ end derived from a rat.

5′-gtcccaaccatgcccaccgatcttccgcctgcttctgaagATGCGGGCCCAG

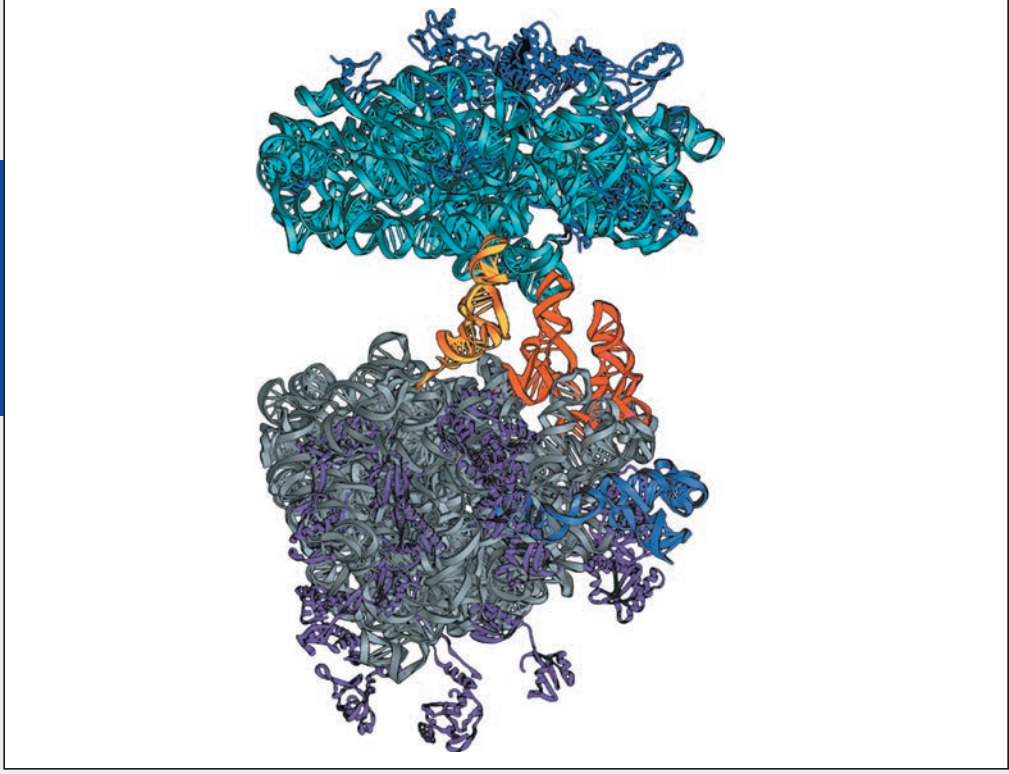
The lowercase letters represent the initial coding portion for the XLαs protein, and the uppercase letters indicate the portion where the ALEX entity is initiated. (For simplicity, and to correspond with the RNA coding dictionary, it is customary to represent the coding (non-template) strand of the DNA segment.)
(a) Convert the coding DNA sequence to the coding RNA sequence.
(b) Locate the initiator codon within the XLαs segment.
(c) Locate the initiator codon within the ALEX segment. Are the two initiator codons in frame?
(d) Provide the amino acid sequence for each coding sequence. In the region of overlap, are the two amino acid sequences the same?
(e) Are there any evolutionary advantages to having the same DNA sequence code for two protein products? Are there any disadvantages?

32. Recent observations indicate that alternative splicing is a common way for eukaryotes to expand their repertoire of gene functions. Studies indicate that approximately 50 percent of human genes exhibit alternative splicing and approximately 15 percent of disease-causing mutations involve aberrant alternative splicing. Different tissues show remarkably different frequencies of alternative splicing, with the brain accounting for approximately 18 percent of such events [Xu et al. (2002). *Nucl. Acids Res.* 30:3754–3766].
(a) Define alternative splicing and speculate on the evolutionary strategy alternative splicing offers to organisms.
(b) Why might some tissues engage in more alternative splicing than others?

33. Isoginkgetin is a cell-permeable chemical isolated from the *Ginkgo biloba* tree that binds to and inhibits snRNPs.
(a) What types of problems would you anticipate in cells treated with isoginkgetin?
(b) Would this be most problematic for *E. coli* cells, yeast cells, or human cells? Why?

14

Translation and Proteins



Crystal structure of a *Thermus thermophilus* 70*S* ribosome, expanded to reveal three bound transfer RNAs, which bridge the two subunits in the fully assembled complex.

CHAPTER CONCEPTS

- The ribonucleotide sequence of messenger RNA (mRNA) reflects genetic information stored in the DNA of genes and corresponds to the amino acid sequences in proteins encoded by those genes.
- The process of translation decodes the information in mRNA, leading to the synthesis of polypeptide chains.
- Translation involves the interactions of mRNA, tRNA, ribosomes, and a variety of translation factors essential to the initiation, elongation, and termination of the polypeptide chain.
- Proteins achieve a three-dimensional conformation that arises from the primary amino acid sequences of the polypeptide chains making up each protein.
- The function of any protein is closely tied to its three-dimensional structure, which can be disrupted by mutation.

In Chapter 13, we established that a genetic code stores information in the form of triplet codons in DNA and that this information is initially expressed, through the process of transcription, as a messenger RNA complementary to the template strand of the DNA helix. However, the final product of gene expression, in the case of protein-coding genes, is a polypeptide chain consisting of a linear series of amino acids whose sequence has been prescribed by the genetic code. In this chapter, we will examine how the information present in mRNA is translated to create polypeptides, which then fold into protein molecules. We will also review the evidence confirming that proteins are the end products of protein-coding genes and discuss briefly the various levels of protein structure, diversity, and function. This information extends our understanding of gene expression and provides an important foundation for interpreting how the mutations that arise in DNA can result in the diverse phenotypic effects observed in organisms.

14.1 Translation of mRNA Depends on Ribosomes and Transfer RNAs

Translation of mRNA is the biological polymerization of amino acids into polypeptide chains. This process, alluded to in our earlier discussion of the genetic code (Chapter 13), occurs only in association with **ribosomes**, which serve as workbenches for polypeptide synthesis. The central question in translation is how triplet codons of mRNA direct specific amino acids into their correct position in the polypeptide. That question was answered once **transfer RNA (tRNA)** was discovered. This class of molecules "adapts" genetic information present as specific triplet codons in mRNA to their corresponding amino acids. Recall that the requirement for some sort of “adaptor” was postulated by Francis Crick in 1957 (Chapter 13).

In association with a ribosome, mRNA presents a triplet codon that calls for a specific amino acid. A specific tRNA molecule contains within its nucleotide sequence three consecutive ribonucleotides complementary to a codon, called the **anticodon**, which can base-pair with the codon. Another region of this tRNA is covalently bound to the codon’s corresponding amino acid.

Base pairing of tRNAs to mRNA holds amino acids in proximity to each other so that a peptide bond can be formed between them. The process occurs over and over as mRNA runs through the ribosome, and amino acids are polymerized into a polypeptide. Before looking more closely at this process, we will first consider the structures of the ribosome and transfer RNA.

Ribosomal Structure

Because of its essential role in the expression of genetic information, the ribosome has been analyzed extensively. A single bacterial cell contains about 10,000 ribosomes, and a eukaryotic cell contains many times more. Electron microscopy revealed that the bacterial ribosome is about 40 nm at its largest diameter and consists of two subunits, one large and one small. Both subunits consist of one or more molecules of **ribosomal RNA (rRNA)** and an array of **ribosomal proteins**. When the two subunits are assembled into a functional ribosome, the structure is sometimes called a **monosome**.

The main differences between bacterial and eukaryotic ribosomes are summarized in **Figure 14.1**. The subunit and rRNA components are most easily isolated and characterized on the basis of their sedimentation behavior when centrifuged in sucrose gradients. The measured rate of migration, abbreviated *S* for *Svedberg coefficient*, reflects a particle’s density, mass, and shape. In bacteria, the monosome is a 70*S* particle, and in eukaryotes it is approximately 80*S*. Note that sedimentation coefficients are not additive. For example, the bacterial 70*S* monosome consists of a 50*S* and a 30*S* subunit, and the eukaryotic 80*S* monosome consists of a 60*S* and a 40*S* subunit.

The larger subunit in bacteria consists of a 23*S* rRNA molecule, a 5*S* rRNA molecule, and 33 ribosomal proteins. In the eukaryotic equivalent, a 28*S* rRNA molecule is accompanied by a 5.8*S* and a 5*S* rRNA molecule and 47 proteins. The smaller bacterial subunits consist of a 16*S* rRNA component and 21 proteins. In the eukaryotic equivalent, an 18*S* rRNA component and about 33 proteins are found. The approximate molecular weights (in daltons, or Da) and the number of nucleotides of the RNA components are also shown in Figure 14.1.

It is now clear that the RNA components of the ribosome perform all-important catalytic functions associated

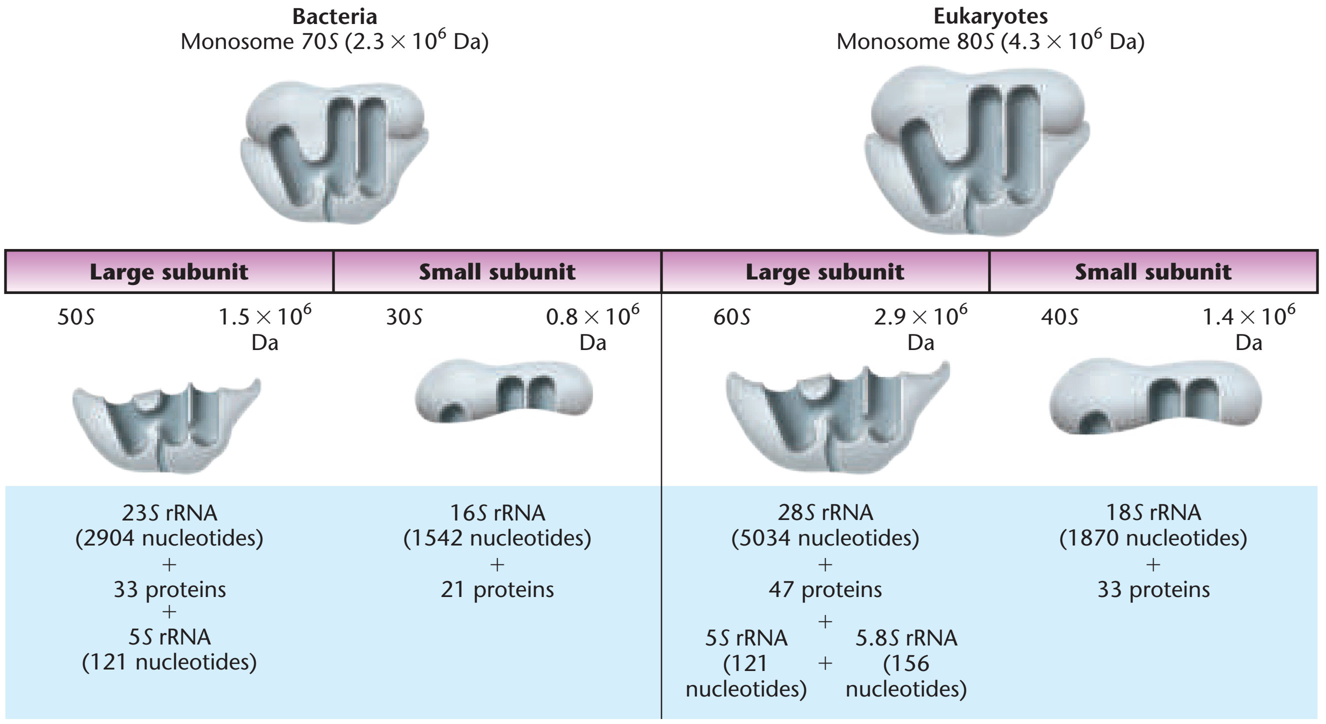

Bacteria
Monosome 70S (2.3×10^6 Da)
Eukaryotes
Monosome 80S (4.3×10^6 Da)
Large subunit
Small subunit
Large subunit
Small subunit
50S
1.5×10^6 Da
30S
0.8×10^6 Da
60S
2.9×10^6 Da
40S
1.4×10^6 Da
23S rRNA (2904 nucleotides) + 33 proteins + 5S rRNA (121 nucleotides)
16S rRNA (1542 nucleotides) + 21 proteins
28S rRNA (5034 nucleotides) + 47 proteins + 5S rRNA (121 nucleotides) + 5.8S rRNA (156 nucleotides)
18S rRNA (1870 nucleotides) + 33 proteins

FIGURE 14.1 A comparison of the components in bacterial and eukaryotic ribosomes. Specific values given are for *E. coli* and human ribosomes.

with translation. The many ribosomal proteins, whose functions were long a mystery, are thought to promote the binding of the various molecules involved in translation and, in general, to fine-tune the process. This conclusion is based on the observation that some of the catalytic functions in ribosomes still occur in experiments involving “ribosomal protein-depleted” ribosomes.

Molecular hybridization studies have established the degree of redundancy of the genes coding for the rRNA components. The *E. coli* genome, for example, contains seven copies of a single sequence that encodes all three components—23*S*, 16*S*, and 5*S*. The initial transcript of each set of these genes produces a 30*S* RNA molecule that is enzymatically cleaved into the smaller components. The coupling of the genetic information encoding these three rRNA components ensures that, following multiple transcription events, equal quantities of all three will be present as ribosomes are assembled.

In eukaryotes, many more copies of a precursor sequence are present. Each copy is initially transcribed into an RNA molecule of about 45*S* that is subsequently processed into 28*S*, 18*S*, and 5.8*S* rRNA components. These species are homologous to the three rRNA components of *E. coli*. In *Drosophila*, approximately 120 copies per haploid genome are present, while in *Xenopus laevis*, more than 500 copies of the larger precursor sequence are present per haploid genome. In mammalian cells, the initial transcript is also 45*S*. The unique 5*S* rRNA component of eukaryotes is not part of this larger transcript. Instead, copies of the gene coding for the 5*S* ribosomal component are distinct and located separately.

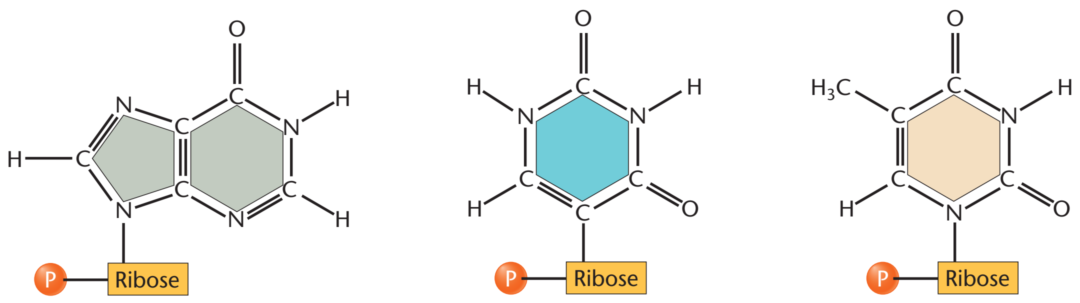
The rRNA genes, called **rDNA**, are part of the middle repetitive DNA fraction of the genome (introduced in Chapter 12) and are present in clusters at various chromosomal sites. Each cluster in eukaryotes consists of **tandem repeats**, with each repeat unit separated by a noncoding **spacer DNA** sequence. In humans, these gene clusters have been localized near the ends of chromosomes 13, 14, 15, 21, and 22. A separate gene cluster encoding 5*S* rRNA has been located on human chromosome 1.

Despite a detailed knowledge of the structure and genetic origin of the ribosomal components, a complete understanding of the function of these components has, to date, eluded geneticists. This is not surprising; the ribosome is the largest and perhaps the most intricate of all cellular structures. For example, the human monosome has a combined molecular weight of 4.3 million Da!

tRNA Structure

Because of their small size and stability in the cell, transfer RNAs (tRNAs) have been investigated extensively and are the best characterized RNA molecules. They are composed of only 75 to 90 nucleotides, displaying a nearly identical structure in bacteria and eukaryotes. In both types of organisms, tRNAs are transcribed from DNA as larger precursors, which are cleaved into mature 4*S* tRNA molecules. Take, for example, tRNA^{Tyr} (the superscript identifies the specific tRNA by the amino acid that binds to it, called its *cognate amino acid*). In *E. coli*, mature tRNA^{Tyr} is composed of 77 nucleotides, yet its precursor contains 126 nucleotides.

In 1965, Robert Holley and his colleagues reported the complete sequence of tRNA^{Ala} isolated from yeast. Of great interest was their finding that a number of nucleotides in the tRNA contain a *modified base* not typically found in mRNA. As illustrated in **Figure 14.2**, each of these nucleotides contains a modification of one of the four nitrogenous bases expected in RNA (G, C, A, and U). Shown are inosinic acid (I), which contains the purine hypoxanthine; ribothymidylic acid (T); and pseudouridylic acid (Ψ). You might want to review the structure of the precursor bases, which are shown in Figure 10.7. These modified structures are created *after* transcription of tRNA, illustrating the more general concept of **posttranscriptional modification**. In this case, an enzymatic reaction modifies the base called for by the genetic code during transcription. These modified bases serve several functions. For example, they confer structural stability and are important for hydrogen bonding between the tRNA and the mRNA being translated.



Inosinic acid (I)

Pseudouridylic acid (Ψ)

Ribothymidylic acid (T)

FIGURE 14.2 Examples of ribonucleotides containing modified nitrogenous bases found in transfer RNA.

Holley's sequence analysis led him to propose the two-dimensional **cloverleaf model of tRNA**. It had been known that tRNA has a characteristic secondary structure created by base pairing. Holley discovered that he could arrange the linear sequence in such a way that several stretches of base pairing would result. His arrangement, with its series of paired stems and unpaired loops, resembled the shape of a cloverleaf. The loops consistently contained modified bases, which did not generally form internal base pairs. Holley's model is shown in **Figure 14.3**.

The triplets GCU, GCC, and GCA specify alanine; therefore, Holley looked for an anticodon sequence complementary to one of these codons in his tRNA^{Ala} molecule. He found it in the form of CGI (read in the 3′ to 5′ direction) in one loop of the cloverleaf. The nitrogenous base I (inosinic acid) can form hydrogen bonds with U, C, or A, the third members of the alanine triplets. Thus, the **anticodon loop** of tRNA^{Ala} was established.

Studies of other tRNA species revealed many constant features. First, at the 3′ end, all tRNAs contain the sequence . . . pCpCpA-3′. At this end of the molecule, the amino acid is covalently joined to the terminal adenosine residue. All tRNAs contain the nucleotide 5′-pG . . . at the other end of the molecule. In addition, the lengths of the analogous stems and loops in tRNA molecules are very similar. Each tRNA examined also contained an anticodon complementary to the known codon for the tRNA's cognate amino acid, and all anticodon loops are present in the same position of the cloverleaf.

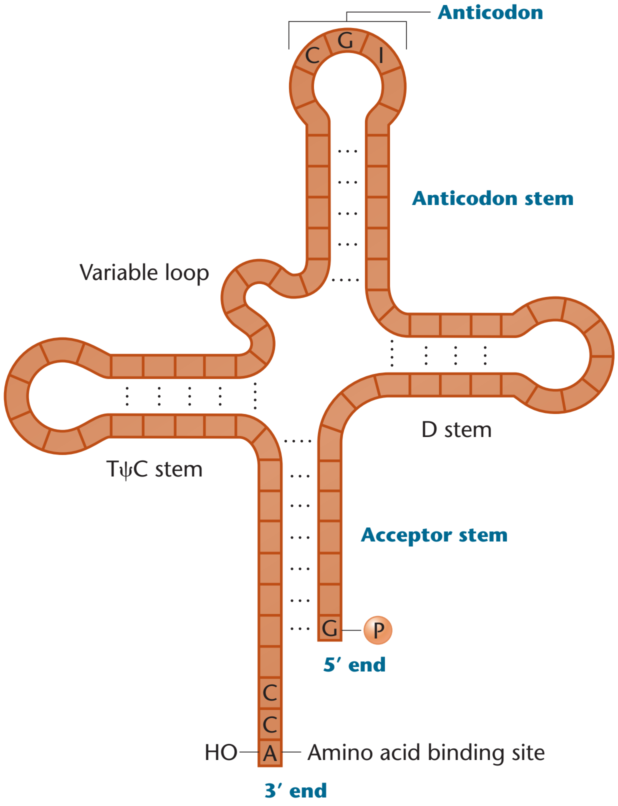


FIGURE 14.3 Holley's two-dimensional cloverleaf model of tRNA^{Ala}. Hydrogen bonds are designated by dots (. . .).

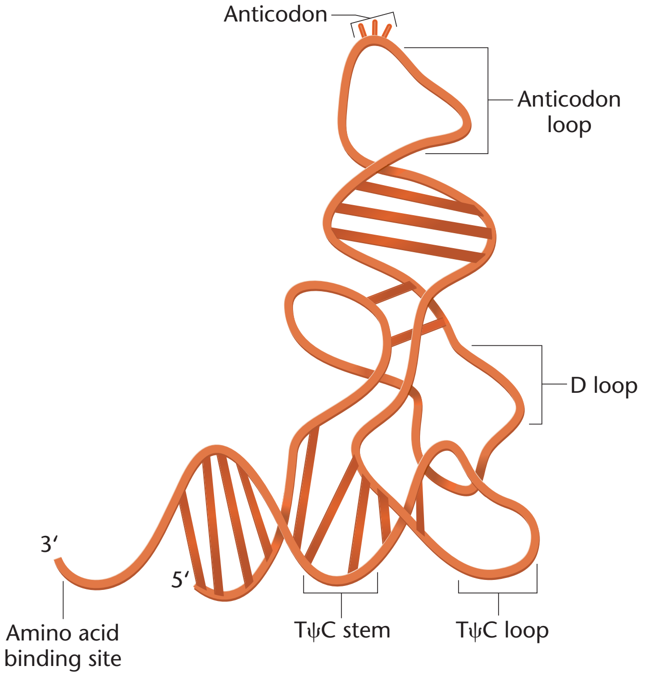


FIGURE 14.4 A three-dimensional model of transfer RNA.

Because the cloverleaf model was predicted strictly on the basis of nucleotide sequence, there was great interest in the three-dimensional structure that would be revealed by X-ray crystallographic examination of tRNA. By 1974, Alexander Rich and his colleagues in the United States, and J. D. Robertus, B. F. C. Clark, Aaron Klug, and their colleagues in England had succeeded in crystallizing tRNA^{Phe} and performing X-ray crystallography at a resolution of 3 Å. At such resolution, the pattern formed by individual nucleotides is discernible.

As a result of these studies, a complete three-dimensional model of a tRNA first became available (**Figure 14.4**). Both the anticodon loop and the 3′-acceptor region (to which the amino acid is covalently linked) were located. Geneticists speculated that the shapes of the intervening loops are recognized by the enzymes responsible for attaching amino acids to tRNAs—a subject to which we now turn our attention.

Charging tRNA

Before translation can proceed, the tRNA molecules must be chemically linked to their respective amino acids. This activation process, called **charging**, or *aminoacylation*, occurs under the direction of enzymes called **aminoacyl tRNA synthetases**. Aminoacyl tRNA synthetases are highly specific enzymes; they recognize only one amino acid and only the tRNAs corresponding to that amino acid, called **isoaccepting tRNAs**. In humans and many other species, there are 20 different aminoacyl tRNA synthetases that catalyze aminoacylation of the 20 different isoacceptor classes of tRNAs. The ability of an aminoacyl tRNA synthetase to recognize specific tRNAs and amino acids is a crucial point if the fidelity of translation is to be maintained. In theory, if the specificity of an aminoacyl tRNA synthetase were changed, it would change the genetic code!

The tRNA charging process is outlined in **Figure 14.5**. In the initial step, the amino acid is converted to an activated form, reacting with ATP to create an **aminoacyladenylic acid**. A covalent linkage is formed between the 5′-end phosphate group of ATP and the carboxyl end of the amino acid. This reaction occurs in association with the synthetase enzyme, forming a complex that then reacts with a specific tRNA molecule. During the next step, the amino acid is transferred to the appropriate tRNA through a high-energy ester bond between the 3′ end of the tRNA and the carboxyl group of the amino acid. The charged tRNA, also referred to as *aminoacyl tRNA*, may then participate directly in protein synthesis.

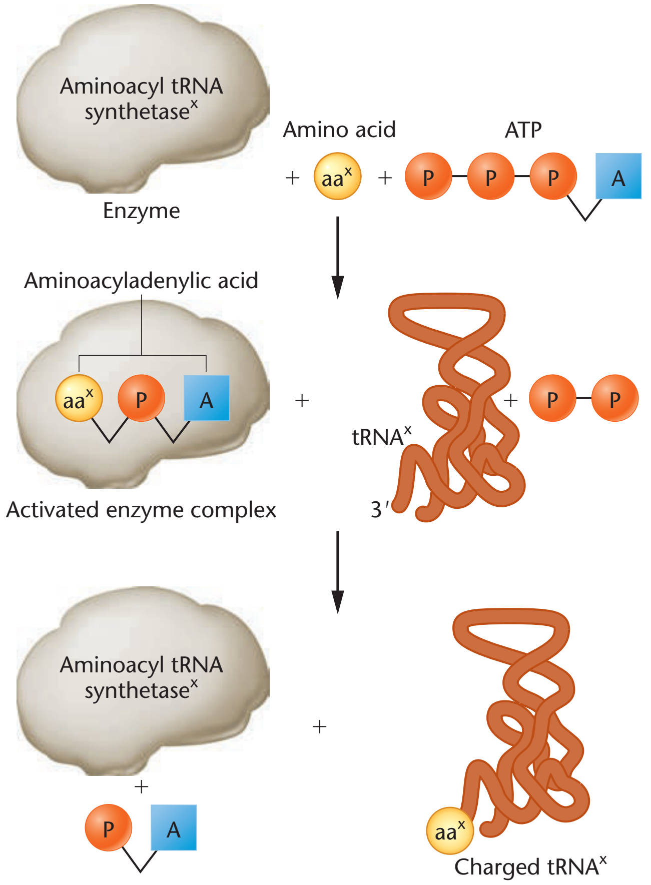


FIGURE 14.5 Steps involved in charging tRNA. The superscript x denotes that only the corresponding specific tRNA and specific aminoacyl tRNA synthetase enzyme are involved in the charging process for each amino acid.

NOW SOLVE THIS

14.1 In 1962, F. Chapeville and others reported an experiment in which they isolated radioactive ^{14}C-cysteinyl-tRNA^{Cys} (charged tRNA^{Cys} + cysteine). They then removed the sulfur group from the cysteine, creating alanyl-tRNA^{Cys} (charged tRNA^{Cys} + alanine). When alanyl-tRNA^{Cys} was added to a synthetic mRNA calling for cysteine, but not alanine, a polypeptide chain was synthesized containing alanine. What can you conclude from this experiment?

- **HINT:** *This problem is concerned with establishing whether the tRNA or the amino acid added to the tRNA during charging is responsible for attracting the charged tRNA to mRNA during translation. The key to its solution is the observation that in this experiment, when the triplet codon in mRNA calls for cysteine, alanine is inserted during translation, even though it is the "incorrect" amino acid.*

For more practice, see Problems 8–10.

14.2 Translation of mRNA Can Be Divided into Three Steps

Much like transcription, the process of translation can best be described by breaking it into discrete phases. We will consider three such phases, each with its own set of illustrations (Figures 14.6, 14.7, and 14.8), but keep in mind that translation is a dynamic, continuous process. As you read the following discussion, keep track of the step-by-step events depicted in the figures. While the core concepts of translation are common for bacterial and eukaryotic cells, the process is simpler in bacteria and is discussed in this section. Many of the protein factors involved in bacterial translation, and their roles, are summarized in **Table 14.1**.

Initiation

Initiation of bacterial translation is depicted in **Figure 14.6**, which at the top, in a box, shows all the individual components involved in the process. Recall that the ribosome serves as a workbench for the translation process. Ribosomes, when they are not involved in translation, are dissociated into their large and small subunits. Note that ribosomes contain three sites, the **aminoacyl (A) site**, the **peptidyl (P) site**, and the **exit (E) site**, the roles of which will soon become apparent. The initiation phase of translation in bacteria requires the association of a small ribosomal subunit, an mRNA molecule, a specific charged initiator tRNA, GTP, Mg^{2+}, and three proteinaceous **initiation factors (IFs)**. In bacteria, the initiation codon of mRNA—AUG—calls for the modified amino acid **N-formylmethionine (fMet)**.

TABLE 14.1 Various Protein Factors Involved during Translation in *E. coli*

Process	Factor	Role
Initiation of translation	IF1	Binds to 30*S* subunit and prevents aminoacyl tRNA from binding to the A site prematurely
	IF2	Binds to the initiator fMet-tRNA and transfers it to the P site of the 30*S*-mRNA complex; releases from complex upon GTP hydrolysis, which is required for 50*S* subunit binding
	IF3	Binds to 30*S* subunit, preventing it from associating with the 50*S* subunit prematurely
Elongation of polypeptide	EF-Tu	Binds GTP; brings aminoacyl tRNA to the A site of the ribosome
	EF-Ts	Regulates EF-Tu activity
	EF-G	Stimulates translocation; GTP-dependent
Termination of translation and release of polypeptide	RF1	Catalyzes release of the polypeptide chain from tRNA and dissociation of the translocation complex; specific for UAA and UAG termination codons
	RF2	Behaves like RF1; specific for UGA and UAA codons
	RF3	Stimulates RF1 and RF2 release

The three initiation factors first bind to the small ribosomal subunit, and this complex in turn binds to mRNA (Step 1). In bacteria, this binding involves a sequence of up to six ribonucleotides (AGGAGG, not shown in Figure 14.6) that *precedes* the initial AUG start codon of mRNA. This sequence—containing only purines and called the **Shine–Dalgarno sequence** —base-pairs with a region of the 16*S* rRNA of the small ribosomal subunit, facilitating initiation.

While IF1 primarily blocks the A site from being bound to a tRNA and IF3 serves to inhibit the small subunit from associating with the large subunit prematurely, IF2 plays a more direct role in initiation. Essentially a GTPase, IF2 interacts with the mRNA and the charged tRNA^{fMet}, stabilizing them in the P site (Step 2). This step "sets" the reading frame so that all subsequent groups of three ribonucleotides are translated accurately. Upon release of IF3, the small subunit (and its associated mRNA and tRNA^{fMet}) then combines with the large ribosomal subunit to create the 70*S* initiation complex. In this process, a molecule of GTP covalently linked to IF2 is hydrolyzed to GDP, causing a conformational change in IF2, and IF1 and IF2 are subsequently released (Step 3).

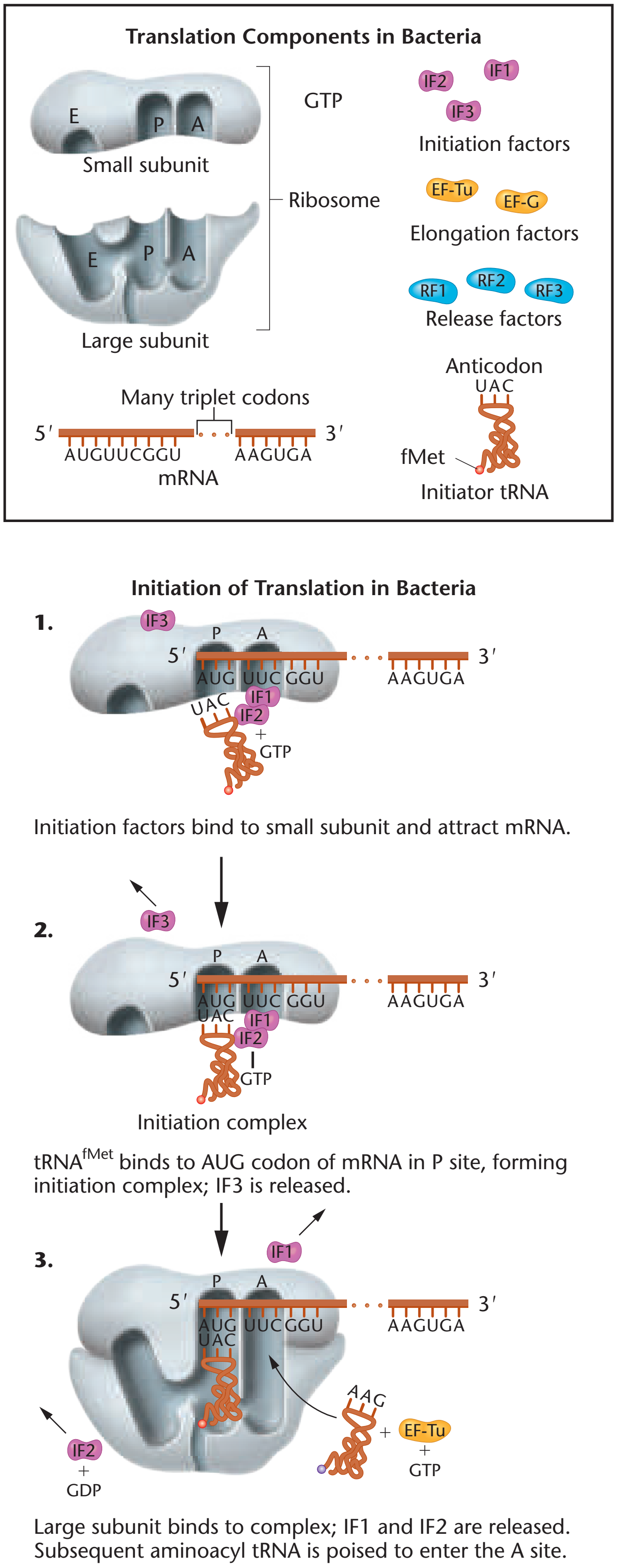


FIGURE 14.6 Initiation of translation in bacteria. (The separate components required for all three phases of translation are depicted in the box at the top of the figure.)

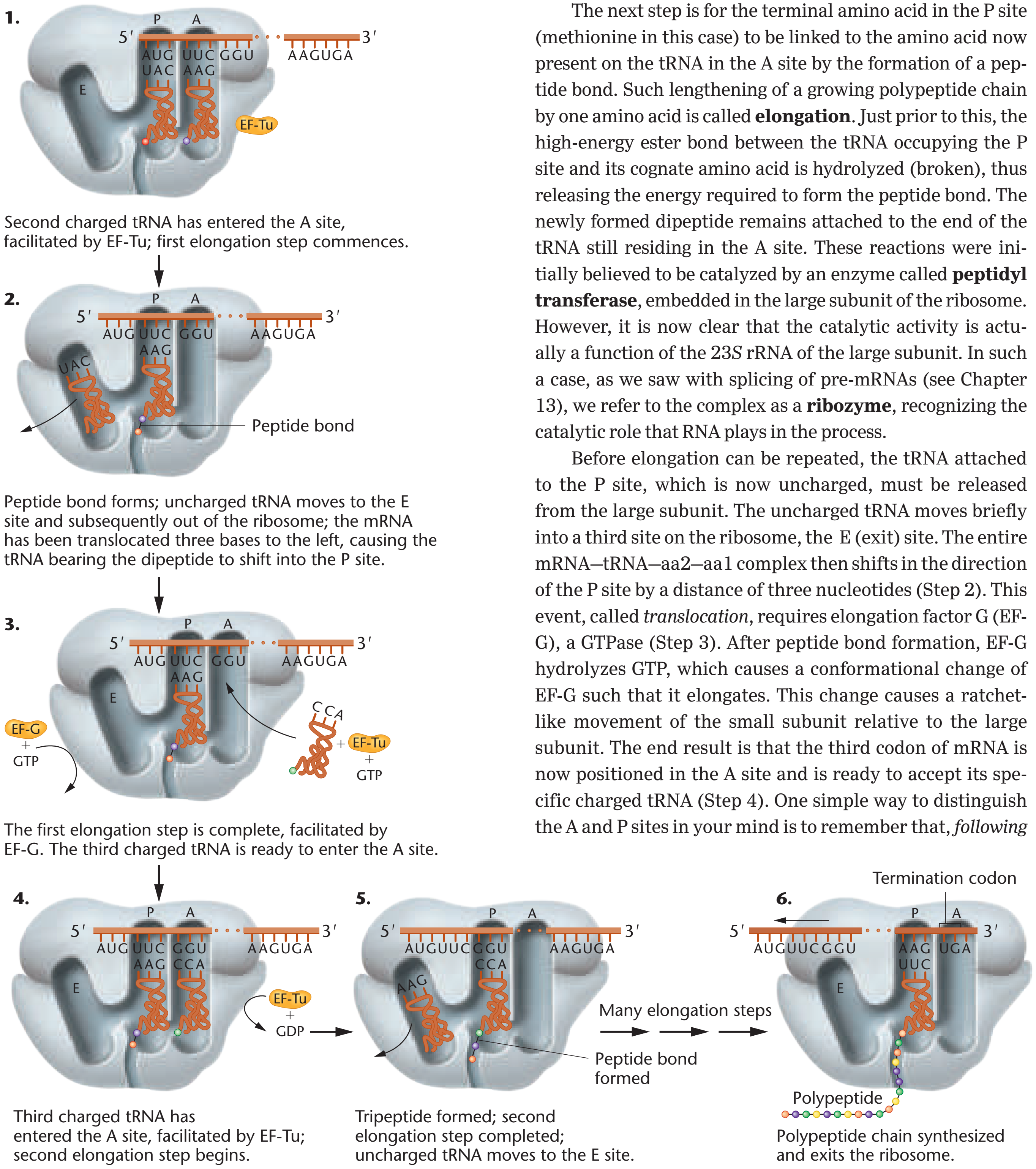
Elongation

The second phase of translation, elongation, is depicted in **Figure 14.7**. As per our prior discussion, the initiation complex is now poised for the insertion into the A site of the second aminoacyl tRNA bearing the amino acid corresponding to the second triplet sequence on the mRNA. Charged tRNAs are transported into the complex by one of the **elongation factors (EFs)**, EF-Tu (Step 1). Like IF2 during initiation, EF-Tu is a GTPase and is bound by a GTP. Hydrolysis of GTP causes a conformational change of EF-Tu such that it releases the bound aminoacyl tRNA.

The next step is for the terminal amino acid in the P site (methionine in this case) to be linked to the amino acid now present on the tRNA in the A site by the formation of a peptide bond. Such lengthening of a growing polypeptide chain by one amino acid is called **elongation**. Just prior to this, the high-energy ester bond between the tRNA occupying the P site and its cognate amino acid is hydrolyzed (broken), thus releasing the energy required to form the peptide bond. The newly formed dipeptide remains attached to the end of the tRNA still residing in the A site. These reactions were initially believed to be catalyzed by an enzyme called **peptidyl transferase**, embedded in the large subunit of the ribosome. However, it is now clear that the catalytic activity is actually a function of the 23*S* rRNA of the large subunit. In such a case, as we saw with splicing of pre-mRNAs (see Chapter 13), we refer to the complex as a **ribozyme**, recognizing the catalytic role that RNA plays in the process.

Before elongation can be repeated, the tRNA attached to the P site, which is now uncharged, must be released from the large subunit. The uncharged tRNA moves briefly into a third site on the ribosome, the E (exit) site. The entire mRNA—tRNA—aa2—aa1 complex then shifts in the direction of the P site by a distance of three nucleotides (Step 2). This event, called *translocation*, requires elongation factor G (EF-G), a GTPase (Step 3). After peptide bond formation, EF-G hydrolyzes GTP, which causes a conformational change of EF-G such that it elongates. This change causes a ratchet-like movement of the small subunit relative to the large subunit. The end result is that the third codon of mRNA is now positioned in the A site and is ready to accept its specific charged tRNA (Step 4). One simple way to distinguish the A and P sites in your mind is to remember that, *following*

Elongation during Translation in Bacteria



1. Second charged tRNA has entered the A site, facilitated by EF-Tu; first elongation step commences.

2. Peptide bond forms; uncharged tRNA moves to the E site and subsequently out of the ribosome; the mRNA has been translocated three bases to the left, causing the tRNA bearing the dipeptide to shift into the P site.

3. The first elongation step is complete, facilitated by EF-G. The third charged tRNA is ready to enter the A site.

4. Third charged tRNA has entered the A site, facilitated by EF-Tu; second elongation step begins.

5. Tripeptide formed; second elongation step completed; uncharged tRNA moves to the E site.

6. Polypeptide chain synthesized and exits the ribosome.

FIGURE 14.7 Elongation of the growing polypeptide chain during translation in bacteria.

translocation, the P site (*P* for peptide) contains a tRNA attached to a peptide chain (a peptidyl tRNA), whereas the A site (*A* for amino acid) contains a charged tRNA with its amino acid attached (an aminoacyl tRNA).

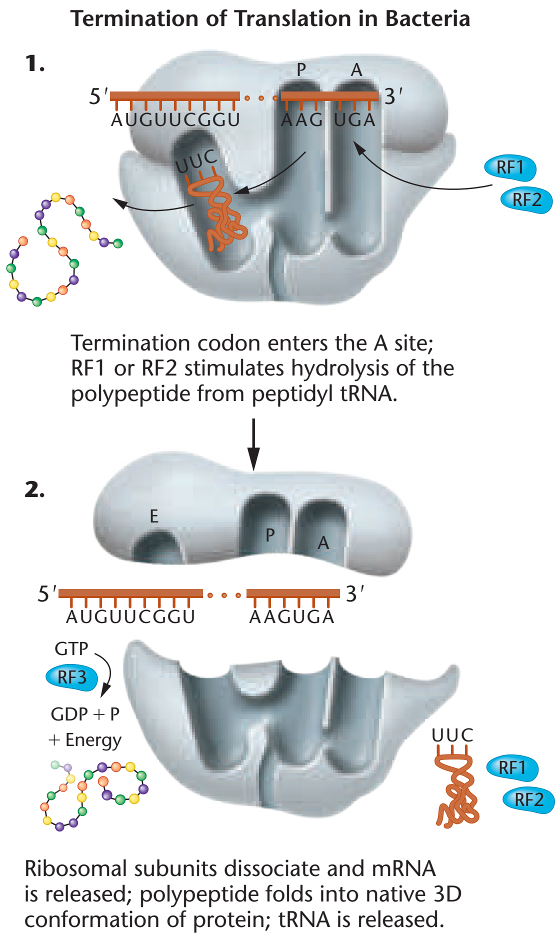
The sequence of elongation and translocation is repeated over and over (Steps 4 and 5). An additional amino acid is added to the growing polypeptide chain each time the mRNA advances by three nucleotides through the ribosome. Once a polypeptide chain of sufficient size is assembled (about 30 amino acids), it begins to emerge from the base of the large subunit, as illustrated in Step 6. The large subunit contains the peptide exit tunnel through which the elongating polypeptide emerges.

As we have seen, the role of the small subunit during elongation is to "decode" the codons in the mRNA, while the role of the large subunit is peptide-bond synthesis. The efficiency of the process is remarkably high: The observed error rate is only about 10^{-4}. At this rate, an incorrect amino acid will occur only once in every 20 polypeptides of an average length of 500 amino acids! In a species such as *E. coli*, elongation proceeds at a rate of about 15 amino acids per second at 37°C.

Termination

Termination, the third phase of translation, is depicted in Figure 14.8. The process is signaled by the presence of any one of three possible triplet codons appearing in the A site: UAG, UAA, or UGA. These codons do not specify an amino acid, nor do they call for a tRNA in the A site. They are called **stop codons**, **termination codons**, or **nonsense codons**. Often, several consecutive stop codons are part of an mRNA. When one such termination stop codon is encountered, the polypeptide, now completed, is still connected to the peptidyl tRNA in the P site, and the A site is empty. The termination codon is not recognized by a tRNA; rather, it is recognized by a **release factor (RF1 or RF2)**, which binds to the A site and stimulates hydrolysis of the polypeptide from the peptidyl tRNA, leading to its release from the translation complex (Step 1). Then, release factor RF3 binds to the ribosome and the tRNA is released from the P site of the ribosome, which then dissociates into its subunits (Step 2).

Interestingly, RF1 is specific to the UAA and UAG termination codons and RF2 is specific to the UAA and the UGA codons. The third release factor, RF3, is a GTPase, and the hydrolysis of a GTP molecule stimulates a conformational change in the ribosome leading to the release of RF1 or RF2 from the stop codon and the tRNA. If a termination codon should appear in the middle of an mRNA molecule as a result of mutation, the same process occurs, and the polypeptide chain is prematurely terminated.



Termination codon enters the A site; RF1 or RF2 stimulates hydrolysis of the polypeptide from peptidyl tRNA.

Ribosomal subunits dissociate and mRNA is released; polypeptide folds into native 3D conformation of protein; tRNA is released.

FIGURE 14.8 Termination of the process of translation in bacteria.

Polyribosomes

As elongation proceeds and the initial portion of an mRNA molecule has passed through the ribosome, this portion of mRNA is free to associate with another small subunit to form a second initiation complex. The process can be repeated several times with a single mRNA and results in what are called **polyribosomes,** or just **polysomes**.

After cells are gently lysed in the laboratory, polyribosomes can be isolated from them and analyzed. The photos in Figure 14.9 show these complexes as seen under the electron microscope. In Figure 14.9(a), you can see the thin lines of mRNA between the individual ribosomes. The micrograph in Figure 14.9(b) is even more remarkable, for it shows the polypeptide chains emerging from the ribosomes during translation. The formation of polysome complexes represents an efficient use of the components available for protein synthesis during a unit of time. It is as if the mRNA is threaded through numerous ribosomes that are side-by-side such that translation is occurring simultaneously in each one, but each subsequent ribosome is a bit behind its neighbor in the amount of mRNA that has been translated.

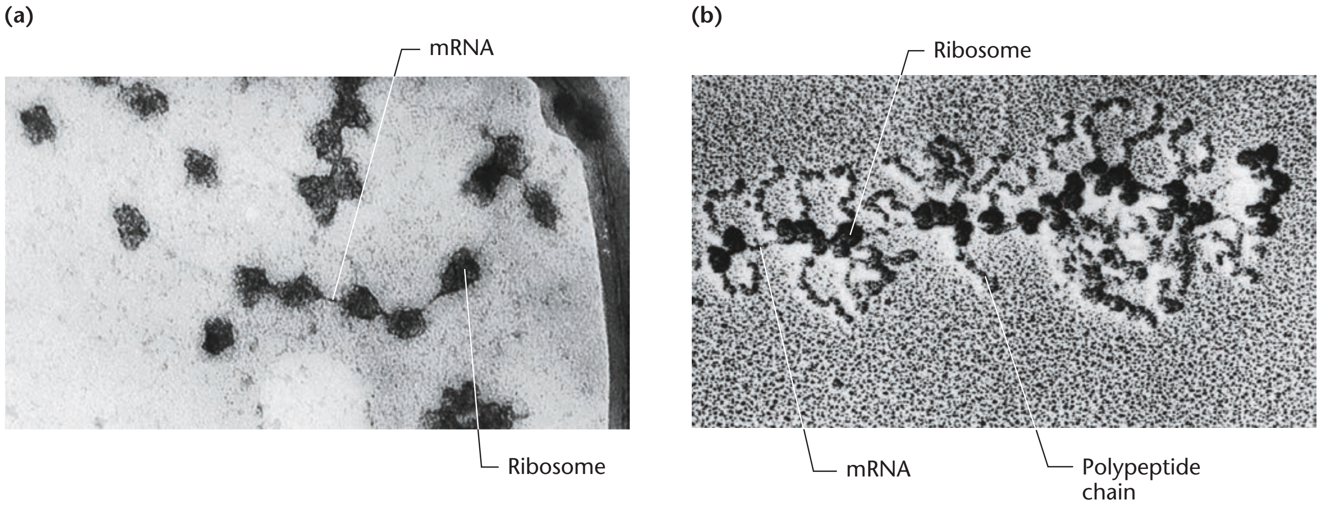


FIGURE 14.9 Polyribosomes as seen under the electron microscope. (a) Sample derived from rabbit reticulocytes engaged in the translation of hemoglobin mRNA. (b) Sample taken from salivary gland cells of the midgefly, *Chironomus thummi*. Note that nascent polypeptide chains are apparent as they emerge from each ribosome. Their length increases as translation proceeds from left to right along the mRNA.

14.3 High-Resolution Studies Have Revealed Many Details about the Functional Bacterial Ribosome

Our knowledge of the process of translation and the structure of the ribosome, as described in the previous sections, was originally based primarily on biochemical and genetic observations, in addition to the visualization of ribosomes under the electron microscope. To confirm and refine this information, the next step was to examine the ribosome at even higher levels of resolution. For example, X-ray diffraction analysis of ribosome crystals was one way to achieve this. However, because of its tremendous size and the complexity of molecular interactions occurring in the functional ribosome, it was extremely difficult to obtain the crystals necessary to perform X-ray diffraction studies. Nevertheless, great strides have been made since 2000. First, the individual ribosomal subunits were crystallized and examined in several laboratories, most prominently that of Venki Ramakrishnan. Then, the crystal structure of the intact 70*S* ribosome, complete with associated mRNA and tRNAs, was examined by Harry Noller and colleagues. In essence, the entire translational complex was seen at the atomic level. Both Ramakrishnan and Noller derived the ribosomes from the bacterium *Thermus thermophilus*. One of the models based on Noller's findings is shown in the opening photograph of this chapter.

Many noteworthy observations have been made from these investigations. For example, the sizes and shapes of the subunits, measured at atomic dimensions are in agreement with earlier estimates based on high-resolution electron microscopy. Furthermore, the shape of the ribosome changes during different functional states, attesting to the dynamic nature of the process of translation. A great deal has also been learned about the prominence and location of the RNA components of the subunits. For example, about one-third of the 16*S* RNA forms a flat projection, referred to as the *platform*, within the smaller 30*S* subunit, that modulates movement of the mRNA–tRNA complex during translocation.

Crystallographic analysis also supports the concept that RNA is the real "player" in the ribosome during translation. The interface between the two subunits, considered to be the location in the ribosome where polymerization of amino acids occurs, is composed almost exclusively of RNA. In contrast, the numerous ribosomal proteins are found mostly on the periphery of the ribosome. These observations confirm what has been predicted on genetic grounds—RNA, not proteins, catalyzes the steps that join amino acids during translation, and thus we may refer to the ribosome as a ribozyme.

Another interesting finding involves the actual location of the various sites predicted to house tRNAs during translation. All three sites (A, P, and E) have been identified in X-ray diffraction studies, and in each case, the RNA of the ribosome makes direct contact with the various loops and domains of the tRNA molecule. This observation supports the hypotheses that had been developed concerning the roles of the different regions of tRNA and helps us understand why the distinctive three-dimensional conformation that is characteristic of all tRNA molecules has been preserved throughout evolution.

Still another noteworthy observation is that the intervals between the A, P, and E sites are at least 20 Å, and perhaps as much as 50 Å, wide, thus defining the atomic distance that the tRNA molecules must shift during each translocation event. This is considered a fairly large distance relative to the size of the tRNAs themselves. Further analysis has led to the identification of RNA–protein bridges existing between the three sites and apparently involved in the translocation events. Other such bridges are present at other key locations and have been related to ribosome function. These observations provide us with a much more complete picture of the dynamic changes that must occur within the ribosome during translation.

A final observation takes us back more than 50 years, to when Francis Crick proposed the **wobble hypothesis** (as introduced in Chapter 13). The Ramakrishnan group has identified the precise location along the 16*S* rRNA of the 30*S* subunit involved in the decoding step that connects mRNA to the proper tRNA. At this location, two particular nucleotides of the 16*S* rRNA actually flip out and probe the codon:anticodon region, and are believed to check for accuracy of base pairing during this interaction. According to the wobble hypothesis, the stringency of this step is high for the first two base pairs but less so for the third (or wobble) base pair.

As our knowledge of the translation process in bacteria has continued to grow, a remarkable study was reported in 2010 by Niels Fischer and colleagues. This research team examined how tRNA is translocated during elongation of the polypeptide chain. Using a unique high-resolution approach—the technique of **resolved single-particle cryo-electron microscopy (cryo-EM)**—the 70*S E. coli* ribosome was captured and examined while in the process of translation at a resolution of 5.5 Å. In this work, over two million images were obtained and computationally analyzed, establishing a temporal snapshot of the trajectories of tRNA during the process of translocation. These results demonstrated that the trajectories are coupled with dynamic conformational changes in the components of the ribosome. Surprisingly, the work revealed that during translation, the ribosome behaves as a complex molecular machine *powered by Brownian motion driven by thermal energy*. That is, the energetic requirement for achieving the various conformational changes essential to translocation are inherent to the ribosome itself.

Numerous questions about ribosome structure and function still remain. In particular, the precise role of the many ribosomal proteins is yet to be clarified. Nevertheless, the models that are emerging based on the work of Noller, Ramakrishnan, Fischer, and their many colleagues provide us with a much better understanding of the mechanism of translation.

14.4 Translation Is More Complex in Eukaryotes

The general features of the model of translation we just discussed were initially derived from investigations of the process in bacteria. Conceptually, the most significant difference between translation in bacteria and eukaryotes is that in bacteria transcription and translation both take place in the cytoplasm and therefore are coupled in bacteria, whereas in eukaryotes these two processes are separated both spatially and temporally. In eukaryotic cells transcription occurs in the nucleus and translation in the cytoplasm. This separation provides multiple opportunities for the regulation of gene expression in eukaryotic cells (a topic we will turn to in Chapters 17 and 18).

Another central difference between bacterial and eukaryotic translation, as we have already seen (Figure 14.1), is that eukaryotes have larger ribosomes composed of a greater number of proteins and RNAs. Interestingly, bacterial and eukaryotic rRNAs share what is called a *core sequence*, but in eukaryotes, rRNAs are lengthened by the addition of *expansion segments* (*ESs*), which are important for ribosome assembly and may also contribute to the regulation and specificity of translation.

The initiation phase is particularly rich in differences between eukaryotes and bacteria. For example, recall that bacterial translation initiation is dependent upon the small subunit pairing with a short sequence upstream of the start codon—the *Shine–Dalgarno sequence*. Eukaryotes lack this sequence. Instead, in a process termed *cap-dependent translation*, initiation in eukaryotes begins with the small subunit associating with the 7-methylguanosine (m^7G) cap located at the 5′ end of eukaryotic mRNAs (Chapter 13). In this context, a specific sequence on bacterial mRNAs and the physical cap structure on eukaryotic mRNAs serve analogous functions.

Recall, too, that bacterial translation initiation requires initiation factors (IF1, IF2, and IF3; Figure 14.6) to attract mRNA to the small subunit and prevent premature association of the large subunit. In eukaryotes, a suite of *eukaryotic* initiation factors (eIFs) carry out these processes. A complex consisting of several eIFs, the initiator tRNA, and the small subunit of the ribosome assemble adjacent to the m^7G cap. The assembly then slides along the mRNA searching for the start codon in a process known as *scanning*. While the bacterial start codon is recognized by an initiator tRNA carrying a N-formylmethionine (fMet), the eukaryotic start codon encodes unformylated methionine. It turns out that a unique transfer RNA ($tRNA_i^{Met}$) is used during eukaryotic initiation, one different from the $tRNA^{Met}$ used for AUG codons after the start. Another eukaryote-specific feature of translation initiation is that many mRNAs contain a purine (A or G) three bases upstream from the AUG initiator codon, which is often followed by a G ($^{A}/_{G}$NN**AUG**G). Named after its discoverer, Marilyn Kozak, this **Kozak sequence** is considered to increase the efficiency of translation initiation in eukaryotes.

Interestingly, the poly-A tail at the 3′ end of eukaryotic mRNAs also plays an important role in translation initiation. Recall from Chapter 13 that **poly-A-binding proteins** bind to the poly-A tail to protect the mRNA from degradation. The poly-A-binding proteins also bind to one of the eukaryotic initiation factors, eIF4G, which in turn binds to eIF4E, also known as the *cap-binding protein*. As this name suggests, eIF4G binds the m^7G cap. The resulting complex is required for translation initiation for many eukaryotic mRNAs. Because the mRNA forms a loop that is closed where the cap and tail are brought together, the process is called **closed-loop translation** (**Figure 14.10**). One possible advantage of closed-loop translation is that the cell will not waste energy translating a partially degraded mRNA lacking either a cap or poly-A tail, features necessary for the closed loop. Another advantage that has been proposed is that the closed-loop structure allows for efficient *ribosome recycling* whereby ribosomes that complete synthesis of one polypeptide can dissociate from the mRNA and then reinitiate translation adjacent to the cap, which is a short distance away in the loop.

Still other differences between translation in bacteria and eukaryotes are noteworthy. Eukaryotic mRNAs are much longer lived than are their bacterial counterparts. Bacterial mRNAs, which lack a cap and poly-A tail, are often translated immediately after transcription and degraded within five minutes. On the other hand eukaryotic mRNAs, with the protection of a cap and poly-A tail, can persist far longer with an average of a 10-hour half-life for mRNAs in cultured human cells. Thus, eukaryotic mRNAs are often available for translation for much longer periods of time.

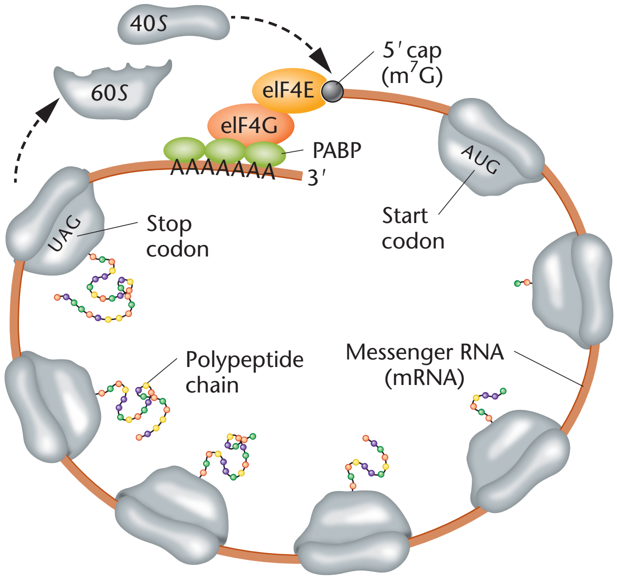


FIGURE 14.10 Eukaryotic closed-loop translation. eIF4E binds to the cap on the mRNA and to a scaffold protein, eIF4G, which binds to poly-A-binding proteins (PABPs) on the poly-A tail of the mRNA. Ribosomes assemble at the cap, scan for the start codon, translate around the loop terminating at a stop codon, and may then reinitiate translation in a process called ribosome recycling.

After translation initiation, proteins similar to those in bacteria guide the elongation and termination of translation in eukaryotes. Many of these **eukaryotic elongation factors (eEFs)** and **eukaryotic release factors (eRFs)** are clearly homologous to their counterparts in bacteria. For example, the role of bacterial EF-Tu, which guides aminoacyl tRNAs into the A site of the ribosome, is fulfilled by eEF-1α. Unlike bacteria, there is a single release factor, eRF1, which recognizes each of the three stop codons in eukaryotes.

We conclude this section by noting that, in 2015, after years of work, the crystal structure of the highly complex 80*S* human ribosome was visualized by Bruno Klaholz and colleagues at the remarkable average resolution of 3.6 Å. This research reveals the interactions between the rRNAs and proteins of the ribosome at an atomic level of detail. In particular, their images show that the interface of the large and small subunits remodels during translation, reflecting a rotational movement of the subunits as the ribosome translocates. Many antibiotics target the bacterial ribosome to inhibit it, but have some negative side effects when used as drugs to fight bacterial infections in humans due to partially inhibiting the human ribosome. This study provides an important model that may assist in reducing the side effects of antibiotics by increasing their specificity for bacterial ribosomes. In addition, this study may enable the design of drugs to slow down the rate of translation of the highly active ribosomes in human cancer cells, thus starving these cells of the protein synthesis on which they are dependent.

14.5 The Initial Insight That Proteins Are Important in Heredity Was Provided by the Study of Inborn Errors of Metabolism

Now, let's consider how we know that proteins are the end products of genetic expression. The first insight into the role of proteins in genetic processes was provided by observations made by Sir Archibald Garrod and William Bateson early in the twentieth century. Garrod became interested in several human disorders that seemed to be inherited. Although he also studied albinism and cystinuria, we will describe his investigation of the disorder

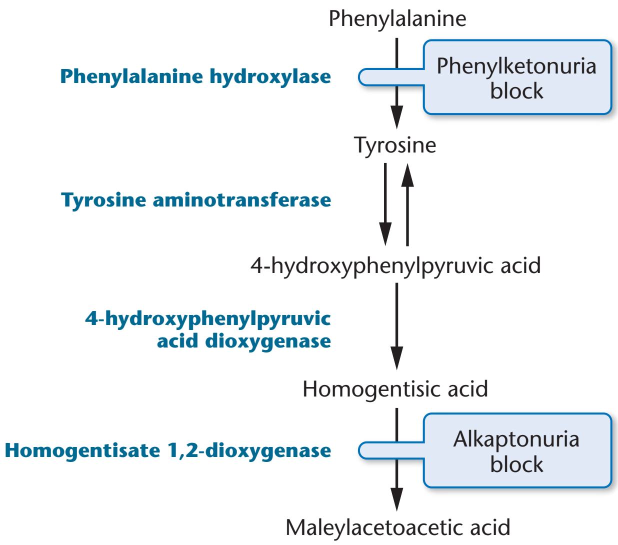

Phenylalanine
Phenylalanine hydroxylase
Phenylketonuria block
Tyrosine
Tyrosine aminotransferase
4-hydroxyphenylpyruvic acid
4-hydroxyphenylpyruvic acid dioxygenase
Homogentisic acid
Homogentisate 1,2-dioxygenase
Alkaptonuria block
Maleylacetoacetic acid

FIGURE 14.11 Metabolic pathway involving phenylalanine and tyrosine. Specific defects in this pathway are implicated in alkaptonuria and phenylketonuria.

alkaptonuria (AKU). Individuals afflicted with this disorder have a disruption in an important metabolic pathway (Figure 14.11). As a result, alkaptonuria patients cannot metabolize a substance called 2,5-dihydroxyphenylacetic acid, also known as homogentisic acid. Homogentisic acid thus accumulates in their cells and tissues and is excreted in the urine. The molecule's oxidation products are black and easily detectable in the diapers of newborns. The unmetabolized products tend to accumulate in cartilaginous areas, causing a darkening of the ears and nose. In joints, this deposition leads to a benign arthritic condition. Alkaptonuria is a rare but not serious disease that persists throughout an individual's life.

Garrod studied alkaptonuria by looking for patterns of inheritance of this benign trait. Eventually he concluded that it was genetic in nature. Of 32 known cases, he ascertained that 19 were confined to seven families, with one family having four affected siblings. In several instances, the parents were unaffected but known to be related as first cousins, and therefore *consanguineous*, a term describing relatives having a common recent ancestor. Parents who are so related have a higher probability than unrelated parents of producing offspring that express recessive traits, because such parents are both more likely to be heterozygous for some of the same recessive traits (see Chapter 25). Garrod concluded that this inherited condition was the result of an alternative mode of metabolism, thus implying that hereditary information controls chemical reactions in the body. While *genes* and *enzymes* were not familiar terms during Garrod's time, he used the corresponding contemporary concepts of *unit factors* and *ferments*. Garrod published his initial observations in 1902.

Only a few geneticists, including Bateson, were familiar with and made reference to Garrod's work. Garrod's ideas fit nicely with Bateson's belief that inherited conditions were caused by the lack of some critical substance. In 1909, Bateson published *Mendel's Principles of Heredity*, in which he linked Garrod's "ferments" with heredity. However, for almost 30 years, most geneticists failed to see the relationship between genes and enzymes. Garrod and Bateson, like Mendel, were ahead of their time.

Phenylketonuria

It is interesting to note that the inherited human metabolic disorder **phenylketonuria (PKU)** results when another reaction in the metabolic pathway shown in Figure 14.11 is blocked. Described first in 1934, this disorder can result in intellectual disability and is transmitted as an autosomal recessive disease. Afflicted individuals are unable to convert the amino acid phenylalanine to the amino acid tyrosine. These molecules differ by only a single hydroxyl group (OH), present in tyrosine but absent in phenylalanine. The reaction is catalyzed by the enzyme **phenylalanine hydroxylase**, which is inactive in affected individuals and active at a level of about 30 percent in heterozygotes. The enzyme functions in the liver. While the normal blood level of phenylalanine is about 1 mg/100 mL, people with phenylketonuria show a level as high as 50 mg/100 mL.

As phenylalanine accumulates, it may be converted to phenylpyruvic acid and, subsequently, to other derivatives. These are less efficiently resorbed by the kidney and tend to spill into the urine more quickly than phenylalanine. Both phenylalanine and its derivatives enter the cerebrospinal fluid, resulting in elevated levels in the brain. The presence of these substances during early development is thought to cause intellectual disability.

Phenylketonuria occurs in approximately 1 in 20,000 births in the United States, where newborns are routinely screened for PKU. When the condition is detected in the analysis of an infant's blood, a strict dietary regimen is instituted in time to prevent intellectual disability. A low-phenylalanine diet can reduce by-products such as phenylpyruvic acid, and the development of abnormalities characterizing the disease can be diminished.

Knowledge of inherited metabolic disorders such as alkaptonuria and phenylketonuria caused a revolution in medical thinking and practice. Human disease, once thought to be solely attributable to the action of invading microorganisms, viruses, or parasites, clearly can have a genetic basis. We know now that hundreds of medical conditions are caused by inborn errors of metabolism resulting from mutant genes. These human biochemical disorders result from defects in the metabolism of all classes of organic biomolecules.

14.6 Studies of *Neurospora* Led to the One-Gene:One-Enzyme Hypothesis

In two separate investigations, George Beadle provided the first convincing experimental evidence that genes are directly responsible for the synthesis of enzymes. The first investigation, begun in 1933 in collaboration with Boris Ephrussi, involved *Drosophila* eye pigments. Together, Beadle and Ephrussi confirmed that mutant genes that altered the eye color of fruit flies could be linked to biochemical errors that, in all likelihood, involved the loss of enzyme function. Encouraged by these findings, Beadle then joined with Edward Tatum to investigate nutritional mutations in the pink bread mold *Neurospora crassa*. This investigation led to the **one-gene:one-enzyme hypothesis**.

Analysis of *Neurospora* Mutants by Beadle and Tatum

In the early 1940s, Beadle and Tatum chose to work with *Neurospora* because much was known about its biochemistry, and mutations could be induced and isolated with relative ease. By inducing mutations, they produced strains that had genetic blocks of reactions essential to the growth of the organism.

Beadle and Tatum knew that the wild-type mold could manufacture nearly every biomolecule necessary for normal development. For example, using rudimentary carbon and nitrogen sources, the organism can synthesize nine water-soluble vitamins, 20 amino acids, numerous carotenoid pigments, and all essential purines and pyrimidines. Beadle and Tatum irradiated asexual conidia (spores) with X rays or UV light to increase the frequency of mutations and then grew the spores on “complete” medium—that is, medium enriched to contain all necessary growth factors (e.g., vitamins and amino acids). Under such growth conditions, a mutant strain unable to grow on minimal medium would be able to grow thanks to the supplements present in the complete medium. All the cultures were then transferred to minimal medium. Any organisms capable of growing on the minimal medium must be able to synthesize all the necessary growth factors themselves, and the researchers could conclude that the cultures from which those organisms came did not contain a nutritional mutation. If no growth occurred, then it was concluded that the culture that had not been able to grow contained a nutritional mutation. These first steps and their results are shown in **Figure 14.12(a)**. The next task was to determine the type of nutritional mutation.

Many thousands of individual mutant spores derived by the same procedure were isolated. To identify the type of mutation the mutant strains were tested on a series of different incomplete media, each containing different groups of supplements [**Figure 14.12(b)**], and subsequently on media containing single vitamins, amino acids, purines, or pyrimidines as supplements, until one specific supplement that permitted growth was found [**Figure 14.12(c)**]. Beadle and Tatum reasoned that the supplement that restored growth was the molecule that the mutant strain could not synthesize. The first mutant strain isolated required vitamin B_6 (pyridoxine) in the medium, and the second one required vitamin B_1 (thiamine). Using the same procedure, Beadle and Tatum eventually isolated and studied hundreds of mutants deficient in the ability to synthesize other vitamins, amino acids, nucleotides, or other substances.

The findings derived from testing more than 80,000 spores convinced Beadle and Tatum that genetics and biochemistry have much in common. It seemed likely that each nutritional mutation caused the loss of the enzymatic activity that facilitates an essential reaction in wild-type organisms. It also appeared that a mutation could be found for nearly any enzymatically controlled reaction. Beadle and Tatum had thus provided sound experimental evidence for the hypothesis that *one gene specifies one enzyme*, an idea alluded to more than 30 years earlier by Garrod and Bateson. With important modifications, this concept was to become another major principle of genetics.

Genes and Enzymes: Analysis of Biochemical Pathways

The one-gene:one-enzyme concept and its attendant research methods have been used over the years to work out many details of metabolism in *Neurospora, Escherichia coli*, and a number of other microorganisms. One of the first metabolic pathways to be investigated in detail was that leading to the synthesis of the amino acid arginine in *Neurospora*. By studying seven mutant strains, each requiring arginine for growth (arg^-), Adrian Srb and Norman Horowitz ascertained a partial biochemical pathway that leads to the synthesis of the amino acid. Their work demonstrates how genetic analysis can be used to establish biochemical information.

Srb and Horowitz tested each mutant strain’s ability to reestablish growth if either citrulline or ornithine, two compounds with close chemical similarity to arginine, was used as a supplement to minimal medium. If either compound was able to substitute for arginine, they reasoned that it must be involved in the biosynthetic pathway of arginine. The researchers found that both molecules could be substituted in one or more strains.

Of the seven mutant strains, four of them (*arg-4* through *arg-7*) grew if supplied with citrulline, ornithine, or arginine. Two of them (*arg-2* and *arg-3*) grew if supplied with citrulline or arginine. One strain (*arg-1*) would

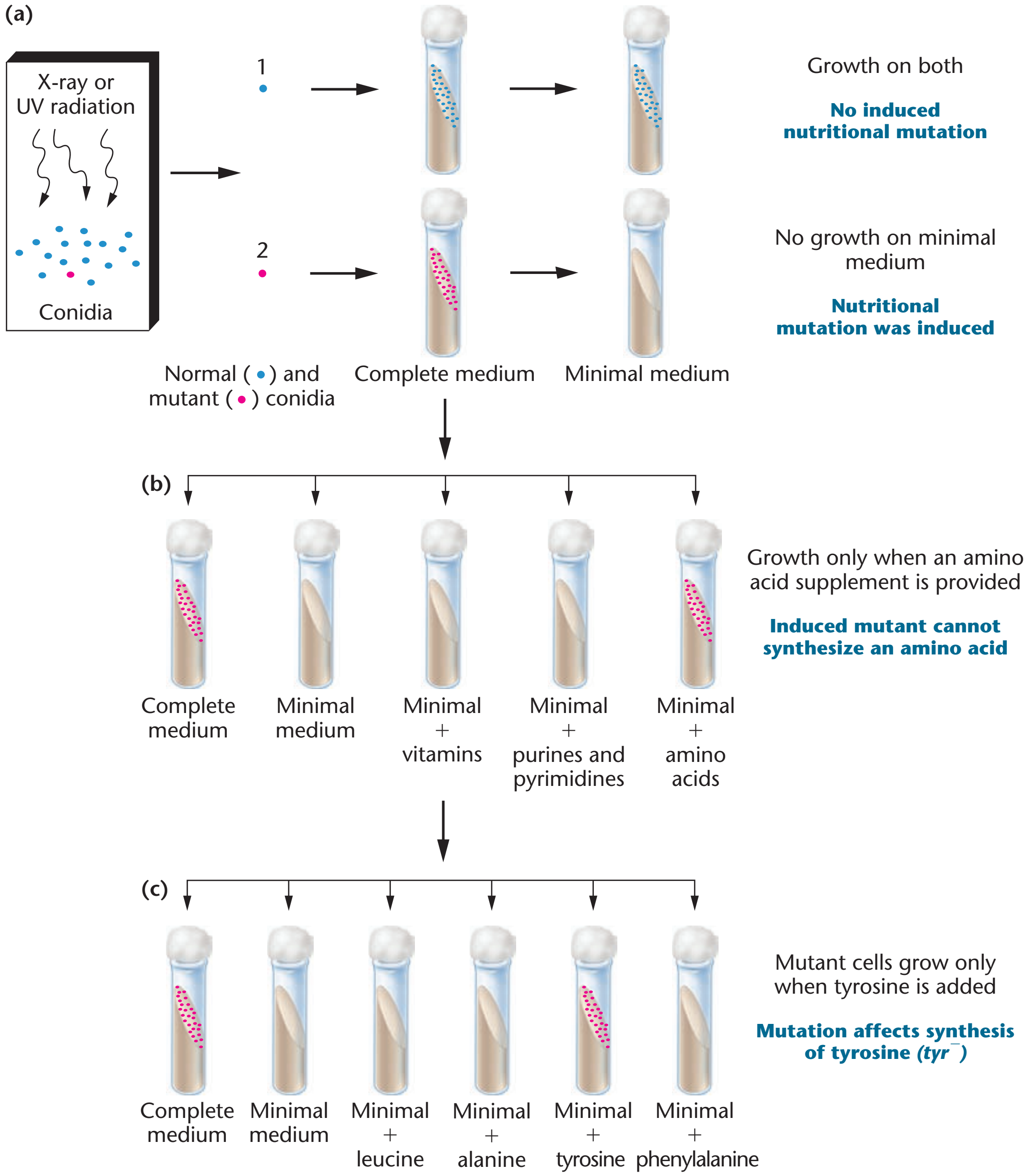

(a)
X-ray or UV radiation
Conidia
1
2
Normal (•) and mutant (•) conidia
Complete medium
Minimal medium
Growth on both
No induced nutritional mutation
No growth on minimal medium
Nutritional mutation was induced
(b)
Complete medium
Minimal medium
Minimal + vitamins
Minimal + purines and pyrimidines
Minimal + amino acids
Growth only when an amino acid supplement is provided
Induced mutant cannot synthesize an amino acid
(c)
Complete medium
Minimal medium
Minimal + leucine
Minimal + alanine
Minimal + tyrosine
Minimal + phenylalanine
Mutant cells grow only when tyrosine is added
Mutation affects synthesis of tyrosine (*tyr*$^-$)

FIGURE 14.12 Induction, isolation, and characterization of a nutritional mutation in *Neurospora crassa*. (a) Most conidia (blue) are not affected by irradiation, but one conidium (shown in red) contains a mutation. (b and c)The precise nature of the mutation is established and found to involve the biosynthesis of tyrosine.

grow only if arginine were supplied; neither citrulline nor ornithine could substitute for it. From these experimental observations (summarized in **Figure 14.13**), the following pathway and metabolic blocks for each mutation were deduced:

$$\text{Precursor} \xrightarrow[\textit{arg-4-7}]{\text{Enzyme A}} \text{Ornithine} \xrightarrow[\textit{arg-2-3}]{\text{Enzyme B}} \text{Citrulline} \xrightarrow[\textit{arg-1}]{\text{Enzyme C}} \text{Arginine}$$

The logic supporting these conclusions is as follows: If mutants *arg-4* through *arg-7* can grow regardless of which of the three molecules is supplied as a supplement to minimal medium, the mutations preventing growth must cause a metabolic block (a defective enzyme "A") that occurs *prior* to the involvement of ornithine, citrulline, or arginine in the pathway. When any one of these three molecules is added, *its presence bypasses the block*. As a result, both citrulline and ornithine appear to be involved in the biosynthesis of arginine. However, the sequence of their participation in the pathway cannot be determined on the basis of these data.

On the other hand, both the *arg-2* and the *arg-3* mutants grow if supplied with citrulline, but not if they are supplied with only ornithine. Therefore, ornithine must be synthesized in the pathway *prior* to the block (a defective enzyme "B"). Presence of ornithine will not overcome

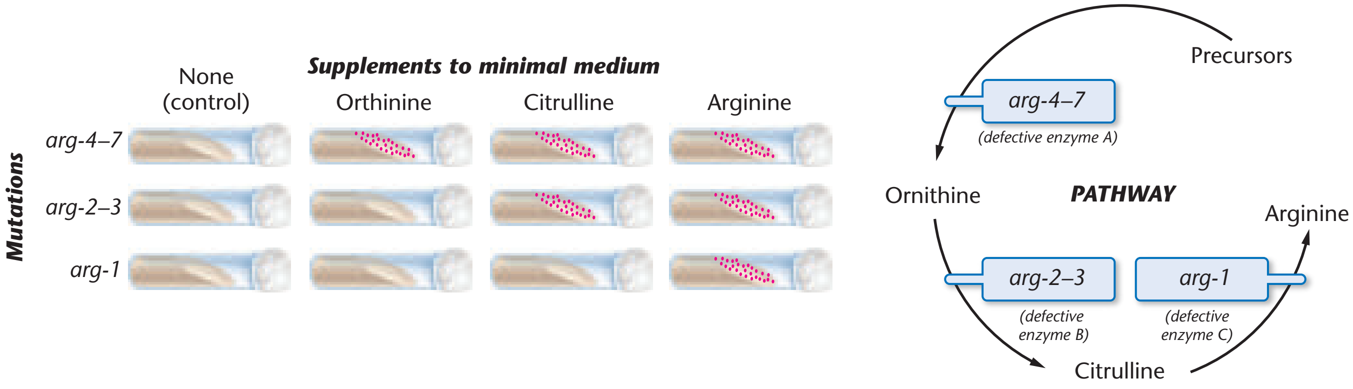


FIGURE 14.13 Srb and Horowitz's experiments on biosynthesis of arginine in *Neurospora crassa* mutants, leading to the formulation of the pathway for arginine production.

the block. Citrulline, however, *does overcome the block*, so it must be synthesized beyond the point of blockage. Therefore, the conversion of ornithine to citrulline represents the correct sequence in the pathway.

Finally, we can conclude that *arg-1* represents a mutation preventing the conversion of citrulline to arginine. Neither ornithine nor citrulline can overcome the metabolic block in this mutation (a defective enzyme "C") because both molecules participate earlier in the pathway.

Taken together, the preceding analysis supports the sequence of biosynthesis shown on the right side of Figure 14.13. Since Srb and Horowitz's experiments in 1944, the detailed pathway has been worked out, and the genes and enzymes controlling each step have been characterized.

NOW SOLVE THIS

14.2 A series of mutations in the bacterium *Salmonella typhimurium* results in the requirement of either tryptophan or some related molecule in order for growth to occur. From the data shown here, suggest a biosynthetic pathway for tryptophan:

	Growth Supplement				
Mutation	Minimal Medium	Anthranilic Acid	Indole Glycerol Phosphate	Indole	Trytophan
trp-8	–	+	+	+	+
trp-2	–	–	+	+	+
trp-3	–	–	–	+	+
trp-1	–	–	–	–	+

■ **HINT:** *This problem asks you to analyze data to establish a biochemical pathway in the bacterium* Salmonella. *The key to its solution is to understand the principles and to apply the same approach that was used to decipher biochemical pathways in* Neurospora.

14.7 Studies of Human Hemoglobin Established That One Gene Encodes One Polypeptide

The one-gene:one-enzyme concept developed in the early 1940s was not immediately accepted by all geneticists. This is not surprising, since it was not yet clear how mutant enzymes could cause variation in the many different kinds of phenotypic traits. For example, *Drosophila* mutants demonstrated altered eye size, wing shape, wing vein pattern, and so on. Plants exhibited mutant varieties of seed texture, height, and fruit size. How an inactive mutant enzyme could result in such phenotypes was puzzling to many geneticists.

Two factors soon modified the one-gene:one-enzyme hypothesis. First, while *nearly all enzymes are proteins, not all proteins are enzymes*. As the study of biochemical genetics proceeded, it became clear that all proteins are specified by the information stored in genes, leading to the more accurate phraseology **one-gene:one-protein hypothesis**. Second, proteins often have a subunit structure consisting of two or more polypeptide chains. This is the basis of the quaternary structure of proteins, which we will discuss in Section 14.8. Because each distinct polypeptide chain is encoded by a separate gene, a more modern statement of Beadle and Tatum's basic principle is **one-gene:one-polypeptide chain hypothesis**. The need for these modifications of the original hypothesis became apparent during the analysis of hemoglobin structure in individuals afflicted with sickle-cell anemia.

Sickle-Cell Anemia

The first direct evidence that genes specify proteins other than enzymes came from the work on mutant hemoglobin molecules derived from humans who have the disorder **sickle-cell anemia**. Affected individuals have erythrocytes

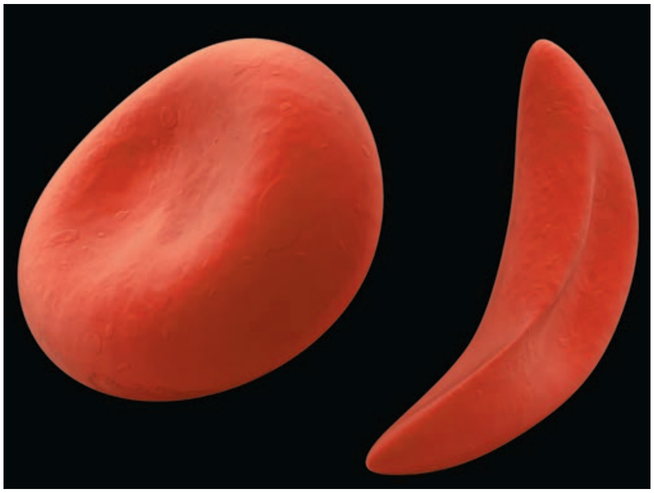


FIGURE 14.14 A comparison of a normal erythrocyte (left) with one derived from a patient with sickle-cell anemia (right).

that, under low oxygen tension, become more rigid, leading to their elongation and increased curvature because of the polymerization of the mutant hemoglobin molecules. The sickle shape of these erythrocytes is in contrast to the biconcave disc shape characteristic in unaffected individuals (Figure 14.14). Those with the disease have attacks in which red blood cells aggregate when oxygen tension is very low, typically in the venous side of capillary systems. As a result, a variety of tissues are deprived of oxygen and suffer severe damage. When this occurs, an individual is said to experience a *sickle-cell crisis* that causes debilitating pain. The kidneys, muscles, joints, brain, gastrointestinal tract, and lungs can be affected. If untreated, a crisis may be fatal.

In addition to undergoing crises, these individuals are anemic because their erythrocytes are destroyed more rapidly than normal red blood cells. Compensatory physiological mechanisms include increased red-cell production by bone marrow and accentuated heart action. These mechanisms lead to abnormal bone size and shape as well as dilation of the heart.

In 1949, James Neel and E. A. Beet each independently demonstrated that the disease is inherited as a Mendelian trait. Pedigree analysis revealed three genotypes and phenotypes controlled by a single pair of alleles, Hb^A and Hb^S. Unaffected and affected individuals result from the homozygous genotypes Hb^A/Hb^A and Hb^S/Hb^S, respectively. The red blood cells of heterozygotes, who exhibit the **sickle-cell trait** but not the disease, undergo much less sickling because more than half of their hemoglobin is normal. Though largely unaffected, such heterozygotes are "carriers" of the defective gene, which is transmitted on average to 50 percent of their offspring.

In the same year, Linus Pauling and coworkers provided the first insight into the molecular basis of sickle-cell anemia. They showed that hemoglobins isolated from diseased and normal individuals differ in their rates of electrophoretic migration. In electrophoresis, charged molecules migrate in an electric field. If the net charge of two molecules is different, the molecules rates of migration will be different. Hence, Pauling and his colleagues concluded that a chemical difference exists between normal and sickle-cell hemoglobin. The two molecules are now designated **HbA** and **HbS**, respectively.

Figure 14.15(a) illustrates the migration pattern of hemoglobin derived from individuals of all three possible genotypes when subjected to **starch gel electrophoresis**. The gel provides the supporting medium for the molecules during migration. In this experiment, samples were placed at a point of origin between the cathode (−) and the anode (+), and an electric current was applied. The migration pattern revealed that all molecules moved toward the anode, indicating a net negative charge. However, HbA migrated farther than HbS, suggesting that its net negative charge was greater. The electrophoretic pattern of hemoglobin derived from individuals who were carriers revealed the presence of both HbA and HbS and confirmed their heterozygous genotype.

Pauling's findings suggested two possibilities. It was known that hemoglobin consists of four nonproteinaceous, iron-containing *heme groups* and a *globin portion* that contains four polypeptide chains. The alteration in net charge in HbS had to be due, theoretically, to a chemical change in one of these components.

Work carried out between 1954 and 1957 by Vernon Ingram resolved this question. He demonstrated that the chemical change occurs in the primary structure of the globin portion of the hemoglobin molecule, specifically in one of the two polypeptides that make up hemoglobin in its quaternary structure. (Human adult hemoglobin contains two identical α chains of 141 amino acids and two identical β chains of 146 amino acids.)

Using a technique called **fingerprinting**, Ingram showed that HbS differs in amino acid composition compared to HbA. The fingerprinting technique involves the enzymatic digestion of the protein into peptide fragments. The mixture is then placed on absorbent paper and exposed to an electric field, where migration occurs based on net charge. The paper is next turned at a right angle to its first exposure and placed in a solvent, in which chromatographic action causes the migration of the peptides in the second direction. The end result is a two-dimensional separation of the peptide fragments into a distinctive pattern of spots, or a "fingerprint." Ingram's work revealed that HbS and HbA differed by only a single peptide fragment [Figure 14.15(b)].

Further analysis then revealed a single amino acid change: Valine was substituted for glutamic acid at the sixth position of the β chain, accounting for the peptide difference [Figure 14.15(c)].

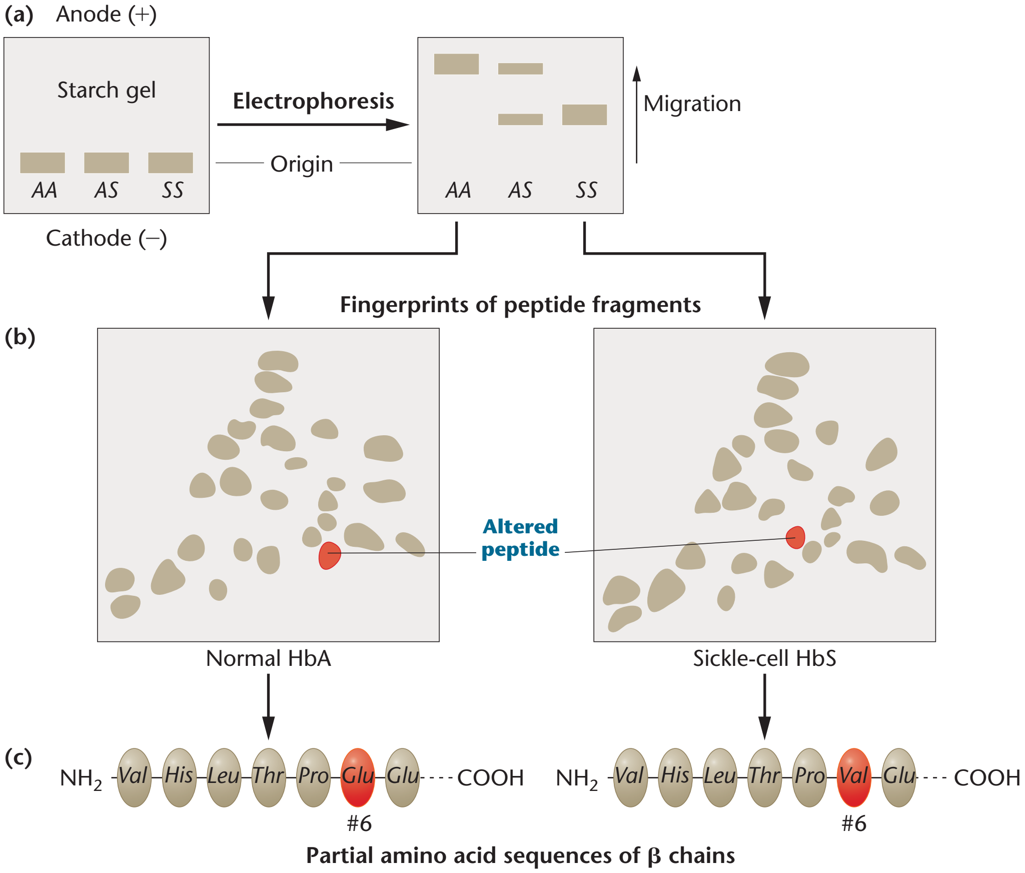

(a)
Anode (+)
Starch gel
AA AS SS
Cathode (–)
Electrophoresis
Origin
AA AS SS
Migration
(b)
Fingerprints of peptide fragments
Altered peptide
Normal HbA
Sickle-cell HbS
(c)
NH2 Val His Leu Thr Pro Glu Glu COOH
#6
NH2 Val His Leu Thr Pro Val Glu COOH
#6
Partial amino acid sequences of β chains

FIGURE 14.15 Investigation of hemoglobin derived from $Hb^A Hb^A$, $Hb^A Hb^S$, and $Hb^S Hb^S$ individuals using electrophoresis, fingerprinting, and amino acid analysis. Hemoglobin from individuals with sickle-cell anemia ($Hb^S Hb^S$): (a) migrates differently in an electrophoretic field; (b) shows an altered peptide in fingerprint analysis; and (c) shows an altered amino acid, valine, at the sixth position in the β chain. During electrophoresis, heterozygotes ($Hb^A Hb^S$)are shown to have both forms of hemoglobin.

The significance of this discovery has been multifaceted. It clearly establishes that a single gene provides the genetic information for a single polypeptide chain. Studies of HbS also demonstrate that a mutation can affect the phenotype by directing a single amino acid substitution. Also, by providing the explanation for sickle-cell anemia, the concept of inherited *molecular disease* was firmly established. Finally, this work led to a thorough study of human hemoglobins, which has provided valuable genetic insights.

In the United States, sickle-cell anemia is found almost exclusively in the African-American population. It affects about 1 in every 365 African-American infants. Currently, about 100,000 individuals are affected. In about 1 of every 170 African-American couples, both partners are heterozygous carriers. In these cases, each of their children has a 25 percent chance of having the disease.

EVOLVING CONCEPT OF THE GENE

In the 1940s, a time when the molecular nature of the gene had yet to be defined, groundbreaking work by Beadle and Tatum provided the first experimental evidence concerning the product of genes, their "one-gene:one-enzyme" hypothesis. This idea received further support and was later modified to indicate that one gene specifies one polypeptide chain or functional RNA. ■

14.8 Variation in Protein Structure Provides the Basis of Biological Diversity

In contrast to nucleic acids, which store and express genetic information, proteins, as end products of genetic expression, are more closely aligned with biological function. It is the variation in biological function that provides the basis of diversity between cell types and between organisms. What is it about proteins that enable them to perform or control enormous numbers of complex and important cellular activities in an organism? As we will see, the secret of the complexity of protein function lies in the incredible structural diversity of proteins.

At the outset of our discussion, we should differentiate between **polypeptides** and **proteins**. Polypeptides, most simply, are precursors of proteins. Thus, as the amino acid polymer is assembled on and then released from the ribosome during translation, it is called a *polypeptide*. Once a polypeptide subsequently folds up and assumes a functional three-dimensional conformation, it is called a *protein*. In many cases, several polypeptides combine during this process to produce an even higher order of protein structure. It is its three-dimensional conformation in space that is essential to a protein's specific function and that distinguishes it from other proteins.

Like nucleic acids, the polypeptide chains comprising proteins are linear, nonbranched polymers. In the vast majority of organisms, there are 20 amino acids that serve as the building blocks of proteins.* Each amino acid has a **carboxyl group**, an **amino group**, and a **radical (R) group** (a side chain that determines the type of amino acid) bound covalently to a **central carbon (C) atom**. **Figure 14.16** shows the 20 R groups that define the 20 amino acids in proteins. The R groups are varied in structure and can be divided into four main classes: (1) **nonpolar (hydrophobic)**, (2) **polar (hydrophilic)**, (3) **positively charged**, and (4) **negatively charged**.

Because polypeptides are often long polymers, and because each unit in the polymer may be any 1 of 20 amino acids, each with unique chemical properties, enormous variation in the molecule's final conformation and chemical activity is possible. For example, if an average polypeptide is composed of 200 amino acids (a molecular weight of about 20,000 Da), 20^{200} different molecules, each with a unique sequence, can be created from the 20 different building blocks.

Around 1900, German chemist Emil Fischer determined the manner in which the amino acids are bonded together. He showed that the amino group (NH^{3-}) of one amino acid can react with the carboxyl group (COO^-) of another amino acid in a dehydration (condensation) reaction, releasing a molecule of H_2O in the process. The resulting covalent bond is called a **peptide bond** (**Figure 14.17**). Two amino acids linked together constitute a *dipeptide*, three a *tripeptide*, and so on. Once ten or more amino acids

*Two other amino acids are exceptions to this rule of 20: **selenocysteine (Sec) and pyrrolysine (Pyl)**, modified versions of cysteine and lysine, respectively. Sec is found in the active sites of a small number of proteins in the three domains of life (archaea, bacteria, and eukaryotes). The insertion of Sec into polypeptides requires a molecular process that recodes UGA codons, which normally function as stop signals, to serve as Sec codons. Similarly, Pyl is inserted into polypeptide chains in response to the stop codon, UAG, when it is present internally in an mRNA. Pyl is found in several methane-loving bacteria and, as discovered more recently, in a bacterium that lives symbiotically under the skin of an annelid worm.

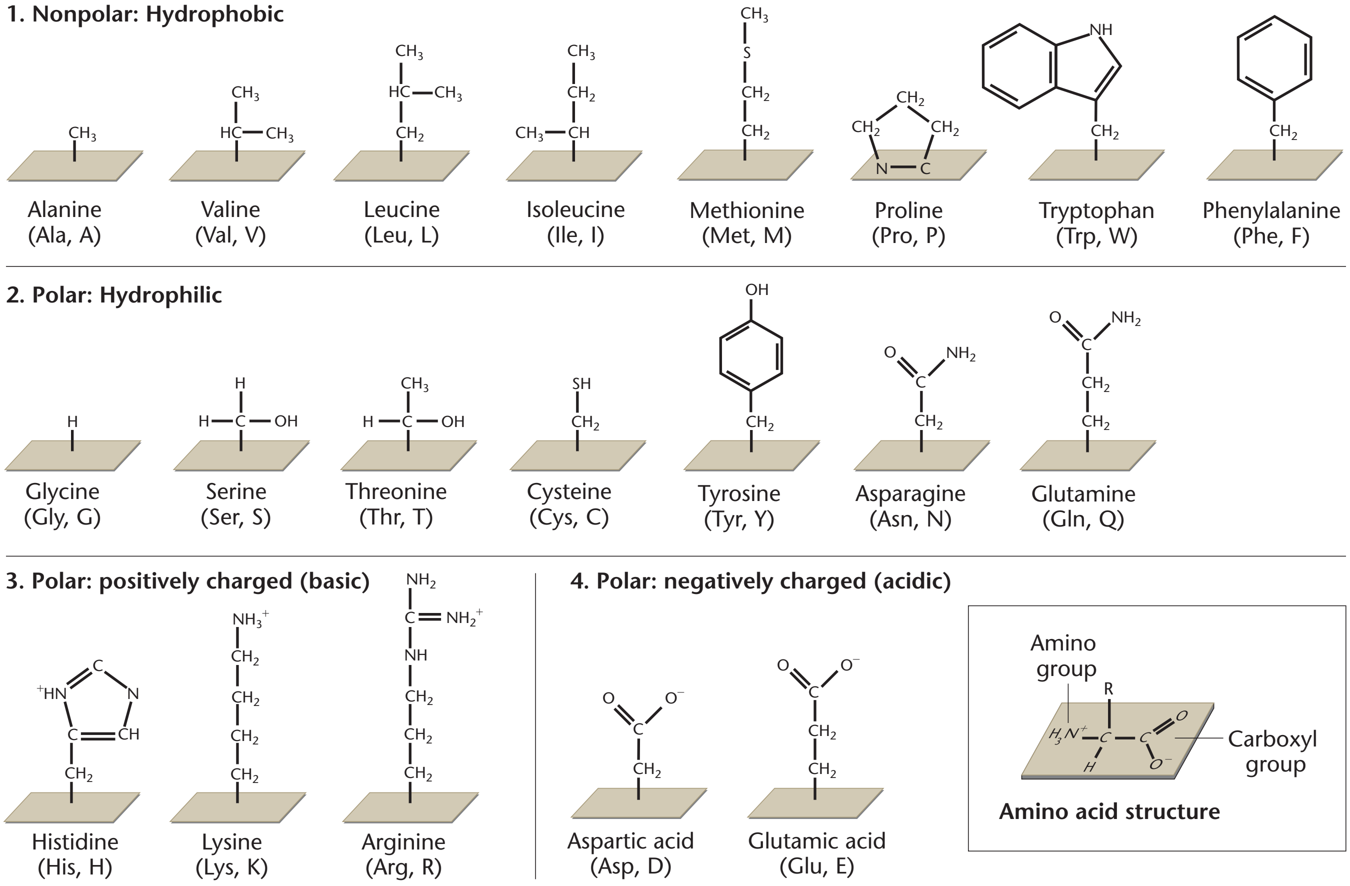


FIGURE 14.16 Chemical structures and designations of the 20 amino acids encoded by living organisms, divided into four major categories. Each amino acid has two abbreviations in universal use; for example, alanine is designated either Ala or A.

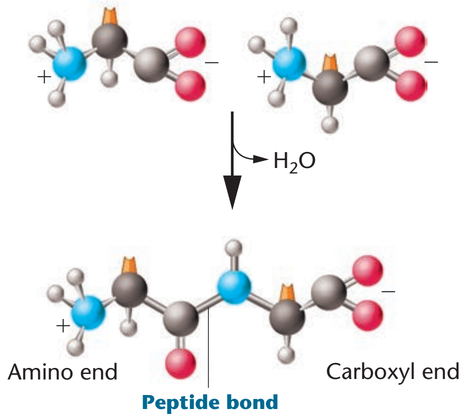


FIGURE 14.17 Peptide bond formation between two amino acids, resulting from a dehydration reaction.

are linked by peptide bonds, the chain is referred to as a *polypeptide*. Generally, no matter how long a polypeptide is, it will have an amino group at one end (the N-terminus) and a carboxyl group at the other end (the C-terminus).

Four levels of protein structure are recognized: primary, secondary, tertiary, and quaternary. The sequence of amino acids in the linear backbone of the polypeptide constitutes its **primary protein structure**. This linear sequence is what is specified by the sequence of deoxyribonucleotides in DNA through an mRNA intermediate. The primary structure of a polypeptide helps determine the specific characteristics of the higher orders of organization as a protein is formed.

Secondary protein structures are certain regular or repeating configurations in space assumed by amino acids lying close to one another in the polypeptide chain. In 1951, Linus Pauling and Robert Corey predicted, on theoretical grounds, an **α helix** as one type of secondary structure. The α-helix model [Figure 14.18(a)] has since been confirmed by X-ray crystallographic studies. It is rodlike and has the greatest possible theoretical stability. The helix is composed of a spiral chain of amino acids stabilized by hydrogen bonds.

The side chains (the R groups) of amino acids extend outward from the helix, and each amino acid residue occupies a distance of 1.5 Å in the length of the helix. There are 3.6 residues per turn. Although left-handed helices are theoretically possible, all α helices seen in proteins are right-handed.

Also in 1951, Pauling and Corey proposed another secondary structure, the **β-pleated sheet**. In this model, a single polypeptide chain folds back on itself, or several chains run in either parallel or antiparallel fashion next to one another. Each such structure is stabilized by hydrogen bonds formed between certain atoms on adjacent chains [Figure 14.18(b)]. In the zigzagging plane formation that results, amino acids in adjacent rows are 3.5 Å apart.

As a general rule, most proteins exhibit a mixture of α-helix and β-pleated sheet structures. Globular proteins, most of which are round in shape and water soluble, usually contain a β-pleated sheet structure at their core, as well as many areas with α helices. The more rigid structural proteins, many of which are water insoluble, rely on more extensive β-pleated sheet regions for their rigidity. For example, **fibroin**, the protein made by the silk moth, depends extensively on this form of secondary structure.

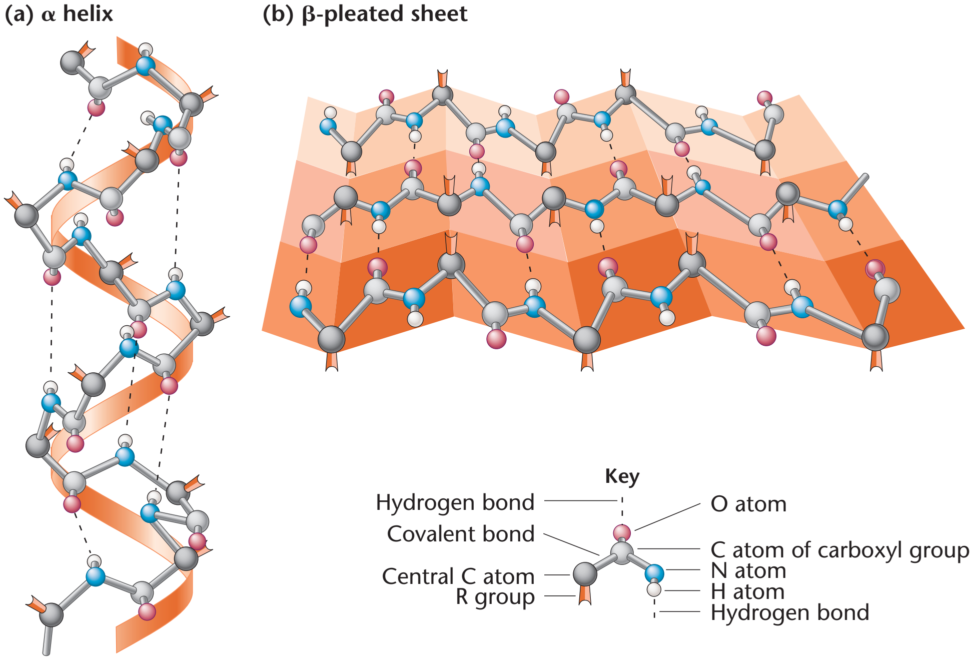


FIGURE 14.18 (a) The right-handed α helix, which represents one form of secondary structure of a polypeptide chain. (b) The β-pleated sheet, an alternative form of secondary structure of polypeptide chains. To maintain clarity, not all atoms are shown.

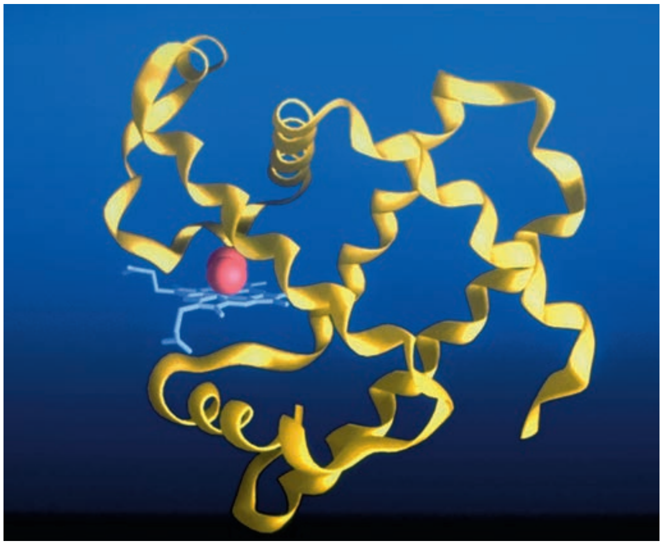


FIGURE 14.19 The tertiary level of protein structure for the respiratory pigment myoglobin. The bound oxygen atom is shown in red.

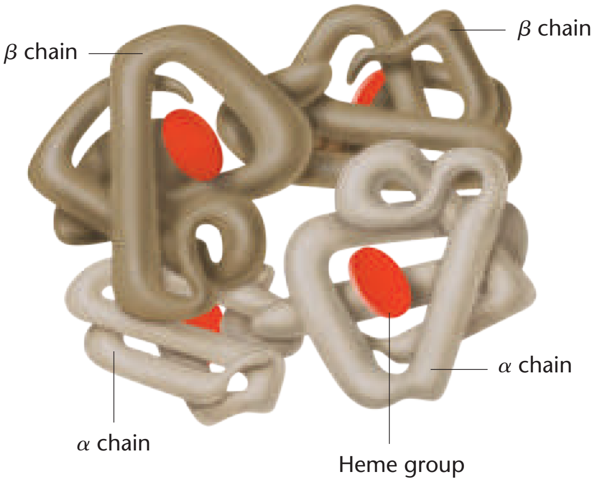


FIGURE 14.20 The quaternary level of protein structure as seen in hemoglobin. Four chains (two αand two β) interact with four heme groups to form the functional molecule.

The secondary structure describes the folding and interactions of amino acids in certain local parts of a polypeptide chain, but the **tertiary protein structure** defines the three-dimensional spatial conformation of the chain as a whole. Each polypeptide twists and turns and loops around itself in a very specific fashion, characteristic of the particular protein. A model of a tertiary structure is shown in **Figure 14.19**.

At the tertiary level of structure, three factors are most important in determining the conformation and in stabilizing the molecule:

1. Covalent disulfide bonds form between closely aligned cysteine residues to form the amino acid dimer cystine.
2. Usually, the polar hydrophilic R groups are located on the surface of the configuration, where they can interact with water.
3. The nonpolar hydrophobic R groups are usually located on the inside of the molecule, where they interact with one another, avoiding interaction with water.

It is important to emphasize that the three-dimensional conformation achieved by any protein is a product of the *primary structure* of the polypeptide. Thus, the genetic code need only specify the sequence of amino acids in order, ultimately, for the final configuration of proteins to be produced. The effects of the three stabilizing factors depend on the location of each amino acid relative to all others in the chain. As folding occurs, the most thermodynamically stable conformation results. The tertiary level of organization is extremely important because the specific function of any protein is directly related to its tertiary structure.

The concept of **quaternary protein structure** applies only to those proteins composed of more than one polypeptide chain and refers to the position of the various chains in relation to one another. This type of protein is *oligomeric*, and each chain in it is a *protomer*, or, less formally, a *subunit*. Protomers have conformations that facilitate their fitting together in a specific complementary fashion. Hemoglobin, an oligomeric protein consisting of four protomers (two α and two β chains), has been studied in great detail. Its quaternary structure is shown in **Figure 14.20**. Many enzymes, including DNA and RNA polymerase, demonstrate quaternary structure.

NOW SOLVE THIS

14.3 HbS results from the substitution of valine for glutamic acid at the number 6 position in the β chain of human hemoglobin. HbC is the result of a change at the same position in the β chain, but in this case lysine replaces glutamic acid. Return to the genetic code table (Figure 13.7) and determine whether single-nucleotide changes can account for these mutations. Then view Figure 14.16 and examine the R groups in the amino acids glutamic acid, valine, and lysine. Describe the chemical differences between the three amino acids. Predict how the changes might alter the structure of the molecule and lead to altered hemoglobin function.

- **Hint:** *This problem asks you to consider the potential impact of several amino acid substitutions that result from mutations in one of the genes encoding one of the chains making up human hemoglobin. The key to its solution is to consider and compare the structure of the three amino acids (glutamic acid, lysine, and valine) and their net charge.*

14.9 Posttranslational Modification Alters the Final Protein Product

Before we turn to a discussion of protein function, it is important to point out that polypeptide chains, like RNA transcripts, are often modified once they have been synthesized. This additional processing is broadly described as **posttranslational modification**. Although many of these alterations are detailed biochemical transformations and beyond the scope of our discussion, you should be aware that they occur and that they are critical to the functional capability of the final protein product. Several examples of posttranslational modification are as follows:

1. *The N-terminal amino acid is often removed or modified.* For example, either the formyl group or the entire formylmethionine residue in bacterial polypeptides is usually removed enzymatically. In eukaryotic polypeptide chains, the amino group of the initial methionine residue is often removed, and the amino group of the N-terminal residue may be modified (acetylated).

2. *Individual amino acid residues are sometimes modified.* For example, phosphates may be added to the hydroxyl groups of certain amino acids, such as tyrosine. Modifications such as these create negatively charged residues that may form an ionic bond with other molecules or change the local conformation of the protein. The process of phosphorylation is extremely important in regulating many cellular activities and is a result of the action of enzymes called *kinases*. At other amino acid residues, methyl groups or acetyl groups may be added enzymatically, which can affect the function of the modified polypeptide chain.

3. *Carbohydrate side chains are sometimes attached.* These are added covalently, producing *glycoproteins*, an important category of cell-surface molecules, such as those specifying the antigens in the ABO blood-type system in humans.

4. *Polypeptide chains may be trimmed.* For example, the insulin gene is first translated into a longer molecule that is enzymatically trimmed to insulin's final 51-amino-acid form.

5. *Target peptides may be removed.* Proteins often function in specific locations of the cell such as the plasma membrane or a particular organelle. During a process known as **protein targeting**, proteins are directed to their appropriate destinations by short internal sequences (3–7 amino acids long) called **target peptides**, which function like postal codes for the cell. For example, a target peptide known as a *signal sequence* directs proteins destined for secretion or embedding in a membrane to the rough endoplasmic reticulum (ER), where they can be inserted into the membrane or translocated into the lumen at the same time they are being synthesized by ribosomes. Similarly, some proteins are directed to mitochondria by a target peptide called the *mitochondrial targeting signal*. Target peptides are often enzymatically cleaved after the protein has been delivered to its proper location.

6. *Folded polypeptide chains are often complexed with prosthetic groups.* The tertiary and quaternary levels of protein structure often include and are dependent on nonproteinaceous elements called *prosthetic groups*, which are commonly vitamins, metals, or metal-containing molecules. The four iron-containing heme groups present in hemoglobin are a good example.

Protein Folding and Misfolding

It was long thought that **protein folding** was a spontaneous process whereby a linear molecule exiting the ribosome achieved a three-dimensional, thermodynamically stable conformation based solely on the combined chemical properties inherent in the amino acid sequence. This indeed is the case for many proteins, and in fact, it is now clear that such a conformational transition begins prior to the polypeptide's exit from the ribosome. Called *cotranslational folding*, this process begins in the peptide exit tunnel during elongation.

However, it has been shown that for other proteins, correct folding is dependent on members of a family of molecules called **chaperones**. Chaperones are themselves proteins (sometimes called *molecular chaperones* or *chaperonins*) that mediate the folding process in one of two ways. Some chaperones simply bind folding polypeptides to exclude the formation of alternative incorrect conformations. Others play a more active role in folding polypeptides by using ATP energy to ensure a proper conformation. While chaperones may bind to the protein in question, like enzymes, they do not become part of the final product.

In eukaryotes, chaperones were initially discovered in *Drosophila*, where they are called **heat-shock proteins** reflecting the fact that protein folding is affected by heat. The heat-shock proteins are expressed in response to high heat to ensure proper protein folding under these conditions. We now know that chaperones are present in all organisms and are even found within mitochondria and chloroplasts.

Even in the presence of chaperones, misfolding may still occur, and one more system of "quality control" exists. Misfolded proteins in bacteria are often digested by ATP-dependent proteases. In eukaryotes, misfolded proteins are often "tagged" with a covalently attached small protein called **ubiquitin**. Enzymes known as **ubiquitin ligases** recognize misfolded proteins and catalyze the attachment of ubiquitin molecules. Once a protein is tagged with several of these residues, it becomes a substrate for the **proteasome**, a large protein complex with protease activity that releases the ubiquitins and degrades the misfolded protein. In addition to eliminating misfolded proteins, ubiquitin-mediated degradation of proteins by the proteasome plays an important role in posttranslational regulation in eukaryotes (see Chapter 18).

Protein folding is a critically important process, not only because misfolded proteins may be nonfunctional, but also because improperly folded proteins can accumulate and be detrimental to cells and the organisms that contain them. For example, a group of transmissible brain disorders in mammals—**scrapie** in sheep, **bovine spongiform encephalopathy (BSE)** (**mad cow disease**) in cattle, and **Creutzfeldt–Jakob disease (CJD)** in humans—are caused by the presence in the brain of aggregates of misfolded proteins called **prions**. The misfolded protein (called PrP^{Sc}) is an altered version of a normal cellular protein (called PrP^{C}) synthesized in neurons and found in the brains of all adult mammals. The difference between PrP^{C} and PrP^{Sc} lies in their secondary protein structures. Normal, noninfectious PrP^{C} folds into a protein with a mainly α-helical structure, whereas infectious PrP^{Sc} folds into a protein with a larger amount of β-pleated sheet structure. When an abnormal PrP^{Sc} molecule contacts a PrP^{C} molecule, the normal protein refolds into the abnormal conformation. The process continues as a chain reaction, with potentially devastating results—the formation of clusters of prions that eventually destroy the brain. Hence, this group of disorders can be considered diseases of secondary protein structure.

Currently, many laboratories are studying protein folding and misfolding, particularly as related to genetics. Numerous inherited human disorders are caused by misfolded proteins that form aggregates. Sickle-cell anemia, discussed earlier in this chapter, is a case in point, where the β chains of hemoglobin are altered as the result of a single amino acid change, causing the molecules to aggregate within erythrocytes, with devastating results. An autosomal dominant inherited form of Creutzfeldt–Jakob disease is known in which the mutation alters the PrP amino acid sequence, leading to prion formation. Various progressive neurodegenerative diseases such as **Huntington disease**, **Alzheimer disease**, and **Parkinson disease** are also linked to the formation of abnormal protein aggregates in the brain. Huntington disease is inherited as an autosomal dominant trait, whereas less clearly defined genetic components are associated with Alzheimer and Parkinson diseases.

14.10 Proteins Perform Many Diverse Roles

The essence of life on Earth resides at the level of diversity of cellular function. While DNA and RNA serve as vehicles for storing and expressing genetic information, proteins are the *means* of cellular function. And it is the capability of cells to assume diverse structures and functions that distinguishes most eukaryotes from simpler organisms such as bacteria. Therefore, an introductory understanding of protein function is critical to a complete view of genetic processes.

Proteins are the most diverse macromolecules found in cells and play many different roles. For example, the respiratory pigments **hemoglobin (Hb)** and **myoglobin** transport oxygen, which is essential for cellular metabolism. **Collagen** and **keratin** are structural proteins associated with the skin, connective tissue, and hair of organisms. **Actin** and **myosin** are contractile proteins, found in abundance in muscle tissue, while **tubulin** is the basis of the function of microtubules in mitotic and meiotic spindles. Still other examples are the **immunoglobulins (IGs)**, which function in the immune system of vertebrates; **transport proteins**, involved in the movement of molecules across membranes; some of the **hormones** and their **receptors**, which regulate various types of chemical activity; **histones**, which bind to DNA in eukaryotic organisms; and **transcription factors** that regulate gene expression.

Nevertheless, the most diverse and extensive group of proteins (in terms of function) are the **enzymes**, to which we have referred throughout this chapter. Enzymes specialize in catalyzing chemical reactions within living cells. Like all catalysts, they increase the rate at which a chemical reaction reaches equilibrium, but they do not alter the end-point of the chemical equilibrium. Their remarkable, highly specific catalytic properties largely determine the metabolic capacity of any cell type and provide the underlying basis of what we refer to as **biochemistry**. The specific functions of many enzymes involved in the genetic and cellular processes of cells are described throughout this text.

Biological catalysis is a process whereby enzymes lower the **energy of activation** for a given reaction. The energy of activation is the increased kinetic energy state that molecules must usually reach before they react with one another. This state can be attained as a result of elevated temperatures, but enzymes allow biological reactions to occur at lower, physiological temperatures. Thus, enzymes make life as we know it possible.

The catalytic properties of an enzyme are determined by the chemical configuration of the molecule's **active site**. This site is associated with a crevice, a cleft, or a pit on the surface of the enzyme that binds reactants, or substrates, and facilitates their interaction. Enzymatically catalyzed reactions control metabolic activities in the cell. Each reaction is either *catabolic* or *anabolic*. **Catabolism** is the degradation of large molecules into smaller, simpler ones accompanied by the release of chemical energy. **Anabolism** is the synthetic phase of metabolism, in which the various components that make up nucleic acids, proteins, lipids, and carbohydrates are built.

14.11 Proteins Often Include More Than One Functional Domain

We conclude this chapter by briefly discussing the important discovery that distinct regions made up of specific amino acid sequences are associated with unique functions in protein molecules. Such sequences, usually between 50 and 300 amino acids, constitute **protein domains** and represent modular portions of the protein that fold into stable, unique conformations independently of the rest of the molecule. Different domains impart different functional capabilities. Some proteins contain only a single domain, while others contain two or more.

The significance of domains resides in the tertiary structures of proteins. Each domain can contain a mixture of secondary structures, including α helices and β-pleated sheets. The unique conformation of a given domain imparts a specific function to the protein. For example, a domain may serve as the catalytic site of an enzyme, or it may impart an ability to bind to a specific ligand. Thus, discussions of proteins may mention *catalytic domains*, *DNA-binding domains*, and so on. In short, a protein must be seen as a collection of structural and functional modules. Obviously, the presence of multiple domains in a single protein increases its versatility and functional complexity.

Exon Shuffling

How do new proteins with novel combinations of protein domains, and thus new functions, evolve? In 1978, Walter Gilbert proposed a model for the evolution of genes encoding new eukaryotic proteins based on a concept he referred to as **exon shuffling**. Gilbert proposed that exons, like protein domains, are also modular and that during evolution exons may have been reshuffled between genes. Most exons are fairly small, averaging about 150 base pairs and encoding about 50 amino acids, consistent with the sizes of many functional domains in proteins. Thus, the shuffling of exons could create new genes encoding proteins with novel combinations of functional domains.

How might exons be shuffled around within the genome? There are two mechanisms important for this process. Transposons, or "jumping genes" (Chapter 15) move from one location to another within the genome through a process known as transposition. Transposons may take other sequences, such as exons, with them when they "jump." Additionally, exons may be shuffled in the genome through errors in DNA recombination such as genetic crossing over at nonhomologous sequences. Since exons are flanked by introns, mechanisms that shuffle exons need not be precise. Changes to flanking intron sequences due to transposition or recombination are not likely to have a negative effect since they are noncoding sequences.

Direct evidence in favor of exon shuffling was presented in 1985 when the human gene encoding the membrane receptor for low-density lipoproteins (LDLs) was isolated and sequenced. The *LDL receptor protein* is essential for the endocytosis of LDL, a carrier of cholesterol in blood plasma, into the cell. Thus, the LDL receptor was predicted to have numerous functional domains, such as those capable of binding specifically to the LDL substrates and regulating endocytosis. In addition, the receptor molecule is modified posttranslationally by the addition of a carbohydrate; a domain must exist that links to this carbohydrate.

Detailed analysis of the gene encoding the LDL receptor supports the concept of exon modules and their shuffling during evolution. For example, most of the various functional domains of the protein are encoded by discrete exons or groups of exons; this underscores the modular relationship of exons and protein domains. In addition, sequence homology between some of the domains of the LDL receptor and other proteins suggests that exons encoding these domains may have been recruited from other genes.

Figure 14.21 shows the relationship between the exons of the LDL receptor gene and the functional domains of the protein. The first exon encodes a signal sequence

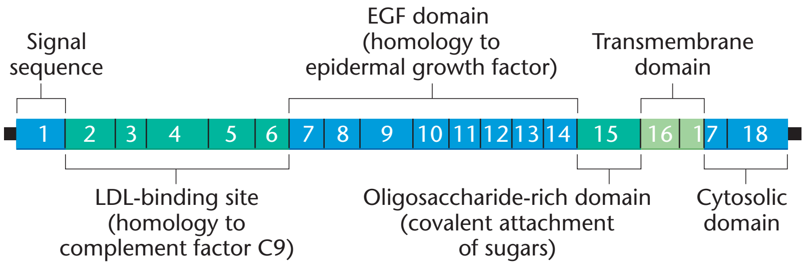
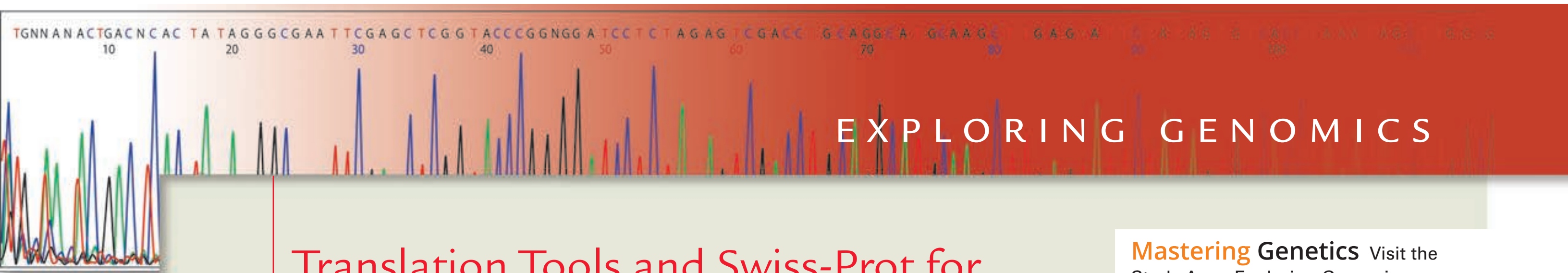


FIGURE 14.21 The 18 exons encoding the LDL receptor protein are organized into five functional domains and one signal sequence.

that is removed from the protein before the LDL receptor becomes part of the membrane. The next five exons, collectively, specify a domain that serves as the binding site for LDL. This domain is homologous with a portion of the gene encoding an immune system protein, *complement factor C9*. The next domain, encoded by eight exons, is called the EGF domain because it bears a striking homology to the peptide-hormone *epidermal growth factor* (*EGF*) (a similar sequence is also found in three blood-clotting proteins). The EGF domain mediates a conformational change of the receptor that is important for the endocytosis of LDL. The 15th exon specifies the oligosaccharide-rich domain; residues within this domain are posttranslationally modified by the addition of carbohydrates. The last two domains include a transmembrane domain that anchors the receptor within the cell membrane and a cytosolic domain that is important for intracellular regulation of LDL endocytosis.

These observations concerning the LDL-receptor exons are fairly compelling in support of the theory of exon shuffling during evolution. Certainly, there is no disagreement concerning the concept of protein domains being responsible for specific molecular interactions.



EXPLORING GENOMICS

Mastering Genetics Visit the Study Area: Exploring Genomics

Translation Tools and Swiss-Prot for Studying Protein Sequences

Many of the databases and bioinformatics programs we have used for Exploring Genomics exercises have focused on manipulating and analyzing DNA and RNA sequences. Scientists working on various aspects of translation and protein structure and function also have a wide range of bioinformatics tools and databases at their disposal.

In this exercise, we will use a bioinformatics portal called **ExPASy (Expert Protein Analysis System)** to translate a segment of a gene into a possible polypeptide. We will then explore other databases to learn more about this polypeptide.

- **Exercise I – Translating a Nucleotide Sequence and Analyzing a Polypeptide**

ExPASy, hosted by the Swiss Institute of Bioinformatics, provides a wealth of resources for studies in proteomics. We will use a program from ExPASy called **Translate Tool** to translate a nucleotide sequence to a polypeptide sequence. Although many other tools are available on the Web for this purpose, ExPASy is one of the more student friendly. Translate Tool allows you to predict a polypeptide sequence from a cloned gene and then look for open reading frames and variations in possible polypeptides.

1. In the MasteringGenetics Study Area, we provide a partial sequence for a human gene based on a complementary DNA (cDNA) sequence. Later in the text (see Chapter 20), you will learn that cDNA sequences are DNA copies complementary to mRNA molecules expressed in a cell. Before you translate this sequence in ExPASy, run a nucleotide–nucleotide search from the NCBI BLAST Web site (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the gene corresponding to this sequence. Refer to the Exploring Genomics exercise in Chapter 10 if you need help with BLAST searches.

2. Access the ExPASy Translate Tool program at http://web.expasy.org/translate/. Copy and paste the cDNA sequence from the Study Area into Translate Tool, and click "Translate Sequence" to generate possible polypeptide sequences encoded by this cDNA.

3. Review the translation results, and then answer the following questions:

   a. Did Translate Tool provide one or multiple possible polypeptide sequences?

   b. If the translation results showed multiple polypeptide sequences, what does this mean? Explain.

   c. Refer to Figure 14.15 in the chapter. Based on part (c) of this figure, which reading frame generated by Translate Tool appears to be correct?

4. ExPASy also provides access to a wealth of information about this polypeptide by connecting to a large number of different databases such as **UniProt Knowledgebase (UniProtKB)**, a protein-sequence database maintained collaboratively by the Swiss Institute for Bioinformatics (SIB), the European Bioinformatics Institute (EMBL-EBI), and a database called the Protein Information Resource (PIR). The UniProt Knowledgebase is widely used by scientists around the world.

(continued)

Exploring Genomics—continued

5. To learn more about the features of this polypeptide, it is best to work with a complete sequence. To obtain a complete sequence, click on the link in the Translate Tool for the reading frame you believe is correct. You will be taken to a results page.

6. To retrieve the amino acid sequence for the entire polypeptide, click "M" for methionine as the first amino acid in the polypeptide and then click the "BLAST submission on ExPASy/SIB" link near the bottom of the page. On the results page, choose UniProt Knowledgebase (UniProtKB) as your database (restricting your search to UniProtKB/Swiss-Prot only), click "Run BLAST," and locate entry P68871; this is an accession number for the protein sequence, similar to the accession numbers assigned to DNA and RNA sequences, and it is the correct match for this sequence. Click on this link to reveal a comprehensive report about the protein. Be sure the identity of this protein agrees with what you discovered in Step 1.

7. On the left side of the UniProtKB report there is a "Display" column with several features listed under it:

   a. Use the "Structure" feature to learn more about this polypeptide.

   b. Under the "Sequence" feature, find the links for Natural Variants to learn more about naturally occurring mutations related to this polypeptide.

   c. Under the "Cross-References" category, find 3D Structure Databases—which presents links to 3D modeling representations showing polypeptide folding arrangements.

   d. Under "Cross-References" locate 2D Gel Databases; use the Swiss-2D PAGE link to view 2D gels of this polypeptide from different tissue samples. (Refer to Figure 21.16 for a representation of 2D gels.) When viewing a 2D gel image with this feature, click on links under "Map Locations" to identify specific spots on a gel that correspond to this polypeptide.

   e. Under "Cross-References" the Family and Domain Databases links will take you to a wealth of information about this polypeptide and related polypeptides and proteins.

   f. Under "Pathology & Biotech" locate the "Chemistry Databases" category and visit the DrugBank link that provides information on drugs that bind to and affect this polypeptide.

CASE STUDY Crippled ribosomes

Diamond–Blackfan anemia (DBA) is a rare, dominant genetic disorder characterized by bone marrow malfunction, birth defects, and a predisposition to certain cancers. Infants with DBA usually develop anemia in the first year of life, have lower than normal production of red blood cells in their bone marrow, and have a high risk of developing leukemia and bone cancer. At the molecular level, DBA is caused by mutations in any one of 10 genes that encode ribosomal proteins. The first-line therapy for DBA is steroid treatment, but more than half of affected children develop resistance to the drugs and in these cases, treatment is halted. DBA can be treated successfully with bone marrow or stem cell transplants from donors with closely matching immune system markers. Transplants from unrelated donors have significant levels of complications and mortality.

1. Given that a faulty ribosomal protein is the culprit and causes DBA, discuss the possible role of normal ribosomal proteins. Why might bone marrow cells be more susceptible to such a mutation than other cells?

2. A couple with a child affected with DBA undergoes *in vitro* fertilization (IVF) and genetic testing of the resulting embryos to ensure that the embryos will not have DBA. However, they also want the embryos screened to ensure that the one implanted can serve as a suitable donor for their existing child. Their plan is to have stem cells from the umbilical cord of the new baby transplanted to their existing child with DBA, thereby curing the condition. What are the ethical pros and cons of this situation?

3. While a stem cell transplant from an unaffected donor is currently the only cure for DBA, genome-editing technologies may one day enable the correction of a mutation in a patient's own bone marrow stem cells. However, what specific information would be needed, beyond a symptom-based diagnosis of DBA, in order to accomplish this?

For related reading, see Penning, G., et al. (2002). Ethical considerations on preimplantation genetic diagnosis for HLA typing to match a future child as a donor of haematopoietic stem cells to a sibling. *Human Reproduction* 17(3):534–538.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Translation is the synthesis of polypeptide chains under the direction of mRNA in association with ribosomes.
2. Translation depends on tRNA molecules that serve as adaptors between triplet codons in mRNA and the corresponding amino acids.
3. Translation occurs in association with ribosomes and, like transcription, is subdivided into the stages of initiation, elongation, and termination and relies on base-pairing affinities between complementary nucleotides.
4. Inherited metabolic disorders are most often due to the loss of enzyme activity resulting from mutations in genes encoding those proteins.
5. Beadle and Tatum's work with nutritional mutations in *Neurospora* led them to propose that one gene encodes one enzyme.
6. Pauling and Ingram's investigations of hemoglobin from patients with sickle-cell anemia led to the modification of the one-gene:one-enzyme hypothesis to indicate that one gene encodes one polypeptide chain.
7. Proteins, the end products of protein-coding genes, demonstrate up to four levels of structural organization that together describe their three-dimensional conformation, which is the basis of each molecule's function.
8. Of the myriad functions performed by proteins, the most influential role belongs to enzymes, which serve as highly specific biological catalysts that play a central role in the production of all classes of molecules in living systems.
9. In eukaryotes, proteins contain one or more functional domains, each prescribed by exon regions interspersed within genes. Specific domains impart specific functional capacities to proteins and appear to have been "shuffled" between genes during evolution.

INSIGHTS AND SOLUTIONS

The growth responses in the following chart were obtained by growing four mutant strains of *Neurospora* on different media, each containing one of four related compounds: A, B, C, and D. None of the mutations grows on minimal medium. Draw all possible conclusions from this data.

	Growth Supplement			
Mutation	A	B	C	D
1	−	−	−	−
2	+	+	−	+
3	+	+	−	−
4	−	+	−	−

Solution: Nothing can be concluded about mutation *1* except that it is lacking some essential factor, perhaps even unrelated to any biochemical pathway in which A, B, C, and D participate. Nor can anything be concluded about compound C. If it is involved in a pathway with the other compounds, it is a product synthesized prior to the synthesis of A, B, and D. We must now analyze these last three compounds and the control of their synthesis by the enzymes encoded by genes *2*, *3*, and *4*. Because product B allows growth in all three cases, it may be considered the "end product"—it bypasses the block in all three instances. Similar reasoning suggests that product A precedes B in the pathway, since A bypasses the block in two of the three steps. Product D precedes B, yielding a partial solution:

$$\text{C(?)} \rightarrow \text{D} \rightarrow \text{A} \rightarrow \text{B}$$

Now let's determine which mutations control which steps. Since mutation *2* can be alleviated by products D, B, and A, it must control a step prior to all three products, perhaps the direct conversion to D, although we cannot be certain. Mutation *3* is alleviated by B and A, so its effect must precede theirs in the pathway. Thus, we will assign it a role controlling the conversion of D to A. Likewise, we can provisionally assign mutation *4* to the conversion of A to B, leading to the following more complete solution.

$$\text{C(?)} \xrightarrow{2(?)} \text{D} \xrightarrow{3} \text{A} \xrightarrow{4} \text{B}$$

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on the translation of mRNA into proteins as well as on protein structure and function. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations in the chapter, what answers would you propose to the following fundamental questions:

   (a) What experimentally derived information led to Holley's proposal of the two-dimensional cloverleaf model of tRNA?

   (b) What experimental information verifies that certain codons in mRNA specify chain termination during translation?

   (c) How do we know, based on studies of *Neurospora* nutritional mutations, that one gene specifies one enzyme?

   (d) On what basis have we concluded that proteins are the end products of genetic expression?

   (e) How do we know that the structure of a protein is intimately related to the function of that protein?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p. 312. These all relate to the translation of genetic information stored in mRNA into proteins and how chemical information in proteins imparts function to those molecules. Write a short essay that discusses the role of ribosomes in the process of translation as it relates to these concepts.

3. Contrast the roles of tRNA and mRNA during translation, and list all enzymes that participate in the transcription and translation process.

4. Francis Crick proposed the “adaptor hypothesis” for the function of tRNA. Why did he choose that description?

5. During translation, what molecule bears the codon? the anticodon?

6. The α chain of eukaryotic hemoglobin is composed of 141 amino acids. What is the minimum number of nucleotides in an mRNA coding for this polypeptide chain?

7. Assuming that each nucleotide in an mRNA is 0.34 nm long, how many triplet codes can simultaneously occupy the space in a ribosome that is 20 nm in diameter?

8. Summarize the steps involved in charging tRNAs with their appropriate amino acids.

9. To carry out its role, each transfer RNA requires at least four specific recognition sites that must be inherent in its tertiary structure. What are they?

10. What are isoaccepting tRNAs? Assuming that there are only 20 different aminoacyl tRNA synthetases but 31 different tRNAs, speculate on parameters that might be used to ensure that each charged tRNA has received the correct amino acid.

11. When a codon in an mRNA with the sequence 5′-UAA-3′ enters the A site of a ribosome, it is not recognized by a tRNA with a complementary anticodon. Why not? What recognizes it instead?

12. Discuss the potential difficulties of designing a diet to alleviate the symptoms of phenylketonuria.

13. Individuals with phenylketonuria cannot convert phenylalanine to tyrosine. Why don’t these individuals exhibit a deficiency of tyrosine?

14. Early detection and adherence to a strict dietary regimen have prevented much of the intellectual disability that used to occur in those with phenylketonuria (PKU). Affected individuals now often lead normal lives and have families. For various reasons, such individuals tend to adhere less rigorously to their diet as they get older. Predict the effect that mothers with PKU who neglect their diets might have on newborns.

15. The synthesis of flower pigments is known to be dependent on enzymatically controlled biosynthetic pathways. For the crosses shown here, postulate the role of mutant genes and their products in producing the observed phenotypes:

    (a) P_1: white strain A × white strain B
    F_1: all purple
    F_2: 9/16 purple: 7/16 white

    (b) P_1: white × pink
    F_1: all purple
    F_2: 9/16 purple: 3/16 pink: 4/16 white

16. The study of biochemical mutants in organisms such as *Neurospora* has demonstrated that some pathways are branched. The data shown in the following table illustrate the branched nature of the pathway resulting in the synthesis of thiamine:

	Growth Supplement			
Mutation	Minimal Medium	Pyrimidine	Thiazole	Thiamine
thi-1	−	−	+	+
thi-2	−	+	−	+
thi-3	−	−	−	+

Why don’t the data support a linear pathway? Can you postulate a pathway for the synthesis of thiamine in *Neurospora*?

17. Explain why the one-gene:one-enzyme concept is not considered totally accurate today.

18. Why is an alteration of electrophoretic mobility interpreted as a change in the primary structure of the protein under study?

19. Using sickle-cell anemia as an example, describe what is meant by a molecular or genetic disease. What are the similarities and dissimilarities between this type of a disorder and a disease caused by an invading microorganism?

20. Contrast the contributions of Pauling and Ingram to our understanding of the genetic basis for sickle-cell anemia.

21. Hemoglobins from two individuals are compared by electrophoresis and by fingerprinting. Electrophoresis reveals no difference in migration, but fingerprinting shows an amino acid difference. How is this possible?

22. HbS results in anemia and resistance to malaria, whereas in those with HbA, the parasite *Plasmodium falciparum* is able to invade red blood cells and cause malaria. Predict whether those with HbC are likely to be anemic and whether they would be resistant to malaria.

23. Several amino acid substitutions in the α and β chains of human hemoglobin are shown in the following table.

Hb Type	Normal Amino Acid	Substituted Amino Acid
Hb Toronto	Ala	Asp (α-5)
HbJ Oxford	Gly	Asp (α-15)
Hb Mexico	Gln	Glu (α-54)
Hb Bethesda	Tyr	His (β-145)
Hb Sydney	Val	Ala (β-67)
HbM Saskatoon	His	Tyr (β-63)

Using the code table (Figure 13.7), determine how many of them can occur as a result of a single-nucleotide change.

24. Define and compare the four levels of protein organization.

25. What are the two common types of protein secondary structure, and how do they differ?

26. How do covalent disulfide bonds, hydrogen bonds with water, and hydrophobic interactions all contribute to a protein’s tertiary structure?

27. List as many different categories of protein functions as you can. Wherever possible, give an example of each category.

28. List three different types of posttranslational modifications that may happen to a protein and the significance of each in the context of protein function.

29. Why are misfolded proteins a potential problem for the eukaryotic cell, and how do cells combat the accumulation of misfolded proteins?

30. How does an enzyme function? Why are enzymes essential for living organisms on Earth?

31. Exon shuffling is a proposal that relates exons in DNA to the repositioning of functional domains in proteins. What evidence exists in support of exon shuffling?

Extra-Spicy Problems

32. Three independently assorting genes (*A, B,* and *C*) are known to control the following biochemical pathway that provides the basis for flower color in a hypothetical plant:

$$\text{Colorless} \xrightarrow{A-} \text{yellow} \xrightarrow{B-} \text{green} \xrightarrow{C-} \text{speckled}$$

Three homozygous recessive mutations are also known, each of which interrupts a different one of these steps. Determine the phenotypic results in the F_1 and F_2 generations resulting from the P_1 crosses of true-breeding plants listed here:
(a) speckled (*AABBCC*) × yellow (*AAbbCC*)
(b) yellow (*AAbbCC*) × green (*AABBcc*)
(c) colorless (*aaBBCC*) × green (*AABBcc*)

33. How would the results vary in cross (a) of Problem 32 if genes *A* and *B* were linked with no crossing over between them? How would the results of cross (a) vary if genes *A* and *B* were linked and 20 map units (mu) apart?

34. Deep in a previously unexplored South American rain forest, a plant species was discovered with true-breeding varieties whose flowers were pink, rose, orange, or purple. A very astute plant geneticist made a single cross, carried to the F_2 generation, as shown:

P_1:	purple × pink
F_1:	all purple
F_2:	27/64 purple
	16/64 pink
	12/64 rose
	9/64 orange

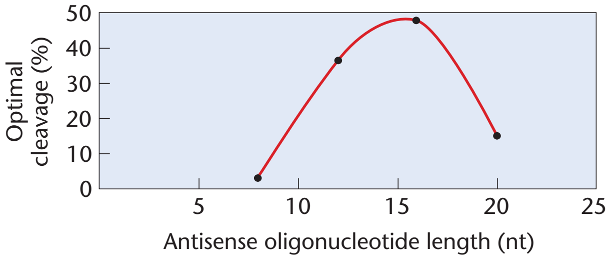
Based solely on these data, he proposed both a mode of inheritance for flower pigmentation and a biochemical pathway for the synthesis of these pigments.

Carefully study the data. Create a hypothesis of your own to explain the mode of inheritance. Then propose a biochemical pathway consistent with your hypothesis. How could you test the hypothesis by making other crosses?

35. Many antibiotics are effective as drugs to fight off bacterial infections because they inhibit protein synthesis in bacterial cells. Using the information provided in the following table that highlights several antibiotics and their mode of action, discuss which phase of translation is inhibited: initiation, elongation, or termination. What other components of the translational machinery could be targeted to inhibit bacterial protein synthesis?

Antibiotic	Action
1. Streptomycin	Binds to 30*S* ribosomal subunit
2. Chloramphenicol	Inhibits peptidyl transferase of 70*S* ribosome
3. Tetracycline	Inhibits binding of charged tRNA to the A site of the ribosome
4. Erythromycin	Binds to free 50*S* particle and prevents formation of 70*S* ribosome
5. Kasugamycin	Inhibits binding of $\text{tRNA}^{\text{fMet}}$
6. Thiostrepton	Prevents translocation by inhibiting EF-G

36. The flow of genetic information from DNA to protein is mediated by messenger RNA. If you introduce short DNA strands (called antisense oligonucleotides) that are complementary to mRNAs, hydrogen bonding may occur and "label" the DNA/RNA hybrid for ribonuclease-H degradation of the RNA. One study [Lloyd et al. (2001). *Nucl. Acids Res.* 29:3664–3673] compared the effect of different-length antisense oligonucleotides upon ribonuclease-H—mediated degradation of tumor necrosis factor (*TNFα*) mRNA. *TNFα* exhibits antitumor and pro-inflammatory activities. The following graph indicates the efficacy of various-sized antisense oligonucleotides in causing ribonuclease-H cleavage.



(a) Describe how antisense oligonucleotides interrupt the flow of genetic information in a cell.
(b) What general conclusion can be drawn from the graph?
(c) What factors other than oligonucleotide length are likely to influence antisense efficacy *in vivo*?

37. Infantile cardiomyopathy is a devastating disorder that is fatal during the first year of life due to defects in the function of heart muscles resulting from mitochondrial dysfunction. A study, performed by Götz et al. [(2011). *Am. J. Hum. Genet.* 88:635–642), identified two different causative mutations in the gene for mitochondrial alanyl-tRNA synthetase (mtAlaRS). One mutation changes a leucine residue at amino acid position 155 to arginine (p.Leu155Arg). The other mutation changes arginine at position 592 to tryptophan (p.Arg592Trp). The mtAlaRS enzyme has an *N*-terminal domain (amino acids 36–481) that catalyzes tRNA aminoacylation and an internal editing domain (amino acids 484–782) that catalyzes deacylation in the case that the tRNA is charged with the wrong amino acid.
(a) Consider the position of the disease causing missense mutations in the *mtAlaRS* gene in the context of the known protein domains of this enzyme. What predictions can you make about how these mutations impair protein synthesis within mitochondria in different ways?
(b) Which mutation would you predict has a more severe impairment of translation in mitochondria, and why?

15

Gene Mutation, DNA Repair, and Transposition

Pigment mutations within an ear of corn, caused by transposition of the *Ds* element.

CHAPTER CONCEPTS

- Mutations comprise any change in the nucleotide sequence of an organism's genome.
- Mutations are a source of genetic variation and provide the raw material for natural selection. They are also the source of genetic damage that contributes to cell death, genetic diseases, and cancer.
- Mutations have a wide range of effects on organisms depending on the type of base-pair alteration, the location of the mutation within the chromosome, and the function of the affected gene product.
- Mutations can occur spontaneously as a result of natural biological and chemical processes, or they can be induced by external factors, such as chemicals or radiation.
- Single-gene mutations cause a wide variety of human diseases.
- Organisms rely on a number of DNA repair mechanisms to counteract mutations. These mechanisms range from proofreading and correction of replication errors to base excision and homologous recombination repair.
- Mutations in genes whose products control DNA repair lead to genome hypermutability, human DNA repair diseases, and cancers.
- Transposable elements may move into and out of chromosomes, causing chromosome breaks and inducing mutations both within coding regions and in gene-regulatory regions.

The ability of DNA molecules to store, replicate, transmit, and decode information is the basis of genetic function. But equally important are the changes that occur to DNA sequences. Without the variation that arises from changes in DNA sequences, there would be no phenotypic variability, no adaptation to environmental changes, and no evolution. Gene mutations are the source of new alleles and are the origin of genetic variation within populations. On the downside, they are also the source of genetic changes that can lead to cell death, genetic diseases, and cancer.

Mutations also provide the basis for genetic analysis. The phenotypic variations resulting from mutations allow geneticists to identify and study the genes responsible for the modified trait. In genetic investigations, mutations act as identifying "markers" for genes so that they can be followed during their transmission from parents to offspring. Without phenotypic variability, classical genetic analysis would be impossible. For example, if all pea plants displayed a uniform phenotype, Mendel would have had no foundation for his research.

We have examined mutations in large regions of chromosomes—chromosomal mutations (see Chapter 8). In contrast, the mutations we will now explore are those occurring primarily in the base-pair sequence of

DNA within and surrounding individual genes—**gene mutations.** We will also describe how the cell defends itself from mutations using various mechanisms of DNA repair.

15.1 Gene Mutations Are Classified in Various Ways

A mutation can be defined as an alteration in the nucleotide sequence of an organism's genome. Any base-pair change in any part of a DNA molecule can be considered a mutation. A mutation may comprise a single base-pair substitution, a deletion or insertion of one or more base pairs, or a major alteration in the structure of a chromosome. The genomes of RNA viruses are made up of single-stranded or double-stranded RNA molecules. These RNA-based genomes are also subject to changes in ribonucleotide sequence that result in mutations. In this chapter, we will restrict our discussion to mutations that occur in DNA genomes.

Mutations may occur within regions of a gene that code for protein or within noncoding regions of a gene such as introns and regulatory sequences, including promoters, enhancers, and splicing signals. Mutations may or may not bring about a detectable change in phenotype. The extent to which a mutation changes the characteristics of an organism depends on which type of cell suffers the mutation and the degree to which the mutation alters the function of a gene product or a gene-regulatory region.

Mutations can occur in somatic cells or within germ cells. Those that occur in germ cells are heritable and are the basis for the transmission of genetic diversity and evolution, as well as genetic diseases. Those that occur in somatic cells are not transmitted to the next generation but may lead to altered cellular function or tumors.

Because of the wide range of types and effects of mutations, geneticists classify mutations according to several different schemes. These organizational schemes are not mutually exclusive. In this section, we outline some of the ways in which gene mutations are classified.

Classification Based on Type of Molecular Change

Geneticists often classify gene mutations in terms of the nucleotide changes that constitute the mutation. A change of one base pair to another in a DNA molecule is known as a **point mutation,** or **base substitution** (see **Figure 15.1**). A change of one nucleotide of a triplet within a protein-coding portion of a gene may result in the creation of a new triplet that codes for a different amino acid in the protein product. If this occurs, the mutation is known as a **missense mutation.** A second possible outcome is that the triplet will be changed into a stop codon, resulting in the termination of translation of the protein. This is known as a **nonsense mutation.** If the point mutation alters a codon but does not result in a change in the amino acid at that position in the protein (due to degeneracy of the genetic code), it can be considered a **silent mutation**.

Because eukaryotic genomes consist of so much more noncoding DNA than coding DNA (see Chapter 12), the vast majority of mutations are likely to occur in noncoding regions. These mutations may be considered **neutral mutations** if they do not affect gene products or gene expression. Most silent mutations, which do not change the amino acid sequence of the encoded protein, can also be considered neutral mutations. However, some silent mutations may alter a DNA sequence that codes for regulatory function, such as an RNA splicing signal, resulting in an altered protein and a discernible phenotype.

You will often see two other terms used to describe base substitutions. If a pyrimidine replaces a pyrimidine or a purine replaces a purine, a **transition** has occurred. If a purine replaces a pyrimidine, or vice versa, a **transversion** has occurred.

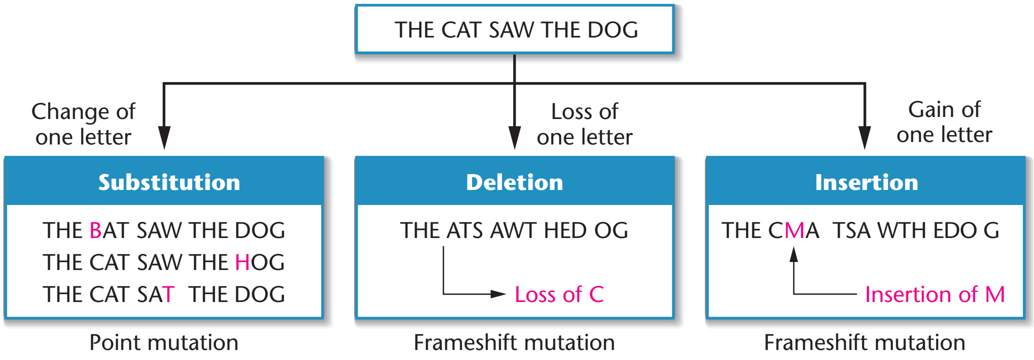


FIGURE 15.1 Analogy showing the effects of substitution, deletion, and insertion of one letter in a sentence composed of three-letter words to demonstrate point and frameshift mutations.

Another type of change is the insertion or deletion of one or more nucleotides at any point within the gene. As illustrated in Figure 15.1, the loss or addition of a single nucleotide causes all of the subsequent three-letter codons to be changed. These are called **frameshift mutations** because the frame of triplet reading during translation is altered. A frameshift mutation will occur when any number of bases are added or deleted, except multiples of three, which would reestablish the initial frame of reading (see Figure 13.2). It is possible that one of the many altered triplets will be UAA, UAG, or UGA, the translation termination codons. When one of these triplets is encountered during translation, polypeptide synthesis is terminated at that point. Obviously, the results of frameshift mutations can be very severe, such as producing a truncated protein or defective enzymes, especially if they occur early in the coding sequence.

Classification Based on Effect on Function

As discussed earlier (see Chapter 4), a **loss-of-function mutation** is one that reduces or eliminates the function of the gene product. Mutations that result in complete loss of function are known as **null mutations.** Any type of mutation, from a point mutation to deletion of the entire gene, may lead to a loss of function.

Most loss-of-function mutations are recessive. A **recessive mutation** results in a wild-type phenotype when present in a diploid organism and the other allele is wild type. In this case, the presence of less than 100 percent of the gene product is sufficient to bring about the wild-type phenotype.

Some loss-of-function mutations can be dominant. A **dominant mutation** results in a mutant phenotype in a diploid organism, even when the wild-type allele is also present. Dominant mutations in diploid organisms can have several different types of effects. A **dominant negative mutation** in one allele may encode a gene product that is inactive and directly interferes with the function of the product of the wild-type allele. For example, this can occur when the nonfunctional gene product binds to the wild-type gene product in a homodimer, inactivating or reducing the activity of the homodimer.

A dominant negative mutation can also result from **haploinsufficiency,** which occurs when one allele is inactivated by mutation, leaving the individual with only one functional copy of a gene. The active allele may be a wild-type copy of the gene but does not produce enough wild-type gene product to bring about a wild-type phenotype. In humans, Marfan syndrome is an example of a disorder caused by haploinsufficiency—in this case as a result of a loss-of-function mutation in one copy of the *fibrillin-1 (FBN1)* gene.

In contrast, a **gain-of-function mutation** codes for a gene product with enhanced, negative, or new functions. This may be due to a change in the amino acid sequence of the protein that confers a new activity, or it may result from a mutation in a regulatory region of the gene, leading to expression of the gene at higher levels or at abnormal times or places. Typically, gain-of-function mutations are dominant.

A **suppressor mutation** is a second mutation that either reverts or relieves the effects of a previous mutation. A suppressor mutation can occur within the same gene that suffered the first mutation (**intragenic mutation**) or elsewhere in the genome (**intergenic mutation**).

An example of an intragenic suppressor mutation is one that reverts a frameshift mutation. For instance, if the original frameshift mutation was caused by a *deletion* of one or two base pairs within a gene's reading frame, the reading frame would be altered, as illustrated in Figure 15.1. However, if a second mutation occurs near the first frameshift mutation, and it involves the *insertion* of one or two base pairs, the original reading frame of the gene may be restored. An incorrect reading frame would still exist for a short distance near the mutations, but the resulting protein would be primarily wild type and could be functional. (The ways in which intragenic suppressor mutations were used experimentally to determine the triplet reading frame of DNA were described in Chapter 13.)

Another example of an intragenic suppressor mutation is one that creates a codon specifying a correct (or similar) amino acid, so as to restore function to a mutated gene product. For instance, if the first mutation changed the sequence 5′-TTA-3′ (which codes for leucine) to 5′-GTA-3′ (which codes for valine), then a second mutation occurring in the valine codon, changing it to 5′-CTA-3′, would restore the codon to one that codes for leucine.

An example of a mutation that would act as an intergenic suppressor mutation would be as follows. A mutated gene may encode a protein whose structure has been altered so that it will not interact with another protein with which it would normally interact. If the gene encoding the second protein acquires a mutation that alters the structure of its gene product in such a way that it can now interact with the first mutant protein, the second mutation would be considered an intergenic suppressor mutation.

Depending on their type and location, mutations can have a wide range of phenotypic effects, from none to severe. Some examples of mutation types based on their phenotypic outcomes are listed in **Table 15.1**.

Classification Based on Location of Mutation

Mutations may be classified according to the cell type or chromosomal locations in which they occur. **Somatic**

TABLE 15.1 Classifications of Mutations by Phenotypic Effects

Classification	Phenotype	Example
Visible	Visible morphological trait	Mendel's pea characteristics
Nutritional	Altered nutritional characteristics	Loss of ability to synthesize an essential amino acid in bacteria
Biochemical	Changes in protein function	Defective hemoglobin leading to sickle-cell anemia in humans
Behavioral	Behavior pattern changes	Brain mutations affecting *Drosophila* mating behaviors
Regulatory	Altered gene expression	Regulatory gene mutations affecting expression of the *lac* operon in *E.coli*
Lethal	Altered organism survival	Tay-Sachs and Huntington disease in humans
Conditional	Phenotype expressed only under certain environmental conditions	Temperature-sensitive mutations affecting coat color in Siamese cats

mutations are those occurring in any cell in the body except germ cells, whereas germ-line mutations occur only in germ cells. **Autosomal mutations** are mutations within genes located on the autosomes, whereas **X-linked** and **Y-linked mutations** are those within genes located on the X or Y chromosome, respectively.

Mutations arising in somatic cells are not transmitted to future generations. When a recessive autosomal mutation occurs in a somatic cell of a diploid organism, it is unlikely to result in a detectable phenotype. The expression of most such mutations is likely to be masked by expression of the wild-type allele within that cell. Somatic mutations will have a greater impact if they are dominant or, in males, if they are X-linked, since such mutations are most likely to be immediately expressed. Similarly, the impact of dominant or X-linked somatic mutations will be more noticeable if they occur early in development, when a small number of undifferentiated cells replicate to give rise to several differentiated tissues or organs. Dominant mutations that occur in cells of adult tissues are often masked by the activity of thousands upon thousands of nonmutant cells in the same tissues that perform the nonmutant function.

Mutations in germ cells are of significance because they may be transmitted to offspring as gametes. They have the potential of being expressed in all cells of an offspring. Inherited dominant autosomal mutations will be expressed phenotypically in the first generation. X-linked recessive mutations arising in the gametes of a female (the **homogametic sex;** having two X chromosomes) may be expressed in male offspring, who are by definition **hemizygous** for the gene mutation because they have one X and one Y chromosome. This will occur provided that the male offspring receives the affected X chromosome. Because of heterozygosity, the occurrence of an autosomal recessive mutation in the gametes of either males or females (even one resulting in a lethal allele) may go unnoticed for many generations, until the resultant allele has become widespread in the population. Usually, the new allele will become evident only when a chance mating brings two copies of it together into the homozygous condition.

NOW SOLVE THIS

15.1 If a point mutation occurs within a human egg cell genome that changes an A to a T, what is the most likely effect of this mutation on the phenotype of an offspring that develops from this mutated egg?

- **Hint:** *This problem asks you to predict the effects of a single base-pair mutation on phenotype. The key to its solution involves an understanding of the organization of the human genome as well as the effects of mutations on coding and noncoding regions of genes and the effects of mutations on development.*

For more practice, see Problems 4–7.

15.2 Mutations Occur Spontaneously and Randomly

When we think about gene mutations, we sometimes assume that genetic errors are predominantly the result of environmental assaults from factors such as toxic chemicals or radiation. Although environmental influences do affect mutation rates, much of the genetic alterations within a genome arise from unknown and endogenous processes. In addition, mutations are not stimulated by evolutionary pressures; they are nondirected. To explore these characteristics of gene mutation, the following subsections describe categories of spontaneous, induced, and random mutations as well as the rates of spontaneous mutations in humans and other organisms.

TABLE 15.2 Spontaneous Mutation Rates at Various Loci in Different Organisms

Organism	Character	Locus	Rate*
Bacteriophage T2	Lysis inhibition	$r \rightarrow r^+$	1×10^{-8}
	Host range	$h^+ \rightarrow h$	4×10^{-9}
Escherichia coli	Lactose fermentation	$lac^- \rightarrow lac^+$	2×10^{-7}
	Streptomycin sensitivity	$str\text{-}d \rightarrow str\text{-}s$	1×10^{-8}
Zea mays	Shrunken seeds	$sh^+ \rightarrow sh^-$	1×10^{-6}
	Purple	$pr^+ \rightarrow pr^-$	1×10^{-5}
Drosophila melanogaster	Yellow body	$y^+ \rightarrow y$	1.2×10^{-6}
	White eye	$w^+ \rightarrow w$	4×10^{-5}
Mus musculus	Piebald coat	$s^+ \rightarrow s$	3×10^{-5}
	Brown coat	$b^+ \rightarrow b$	8.5×10^{-4}

* Rates are expressed per gene replication (T2), per cell division (*Escherichia coli*), or per gamete per generation (*Zea mays*, *Drosophila melanogaster*, and *Mus musculus*).

Spontaneous and Induced Mutations

Mutations can be classified as either spontaneous or induced, although these two categories overlap to some degree. **Spontaneous mutations** are changes in the nucleotide sequence of genes that appear to occur naturally. No specific agents are associated with their occurrence. Many of these mutations arise as a result of normal biological or chemical processes in the organism that alter the structure of nitrogenous bases. Often, spontaneous mutations occur during the enzymatic process of DNA replication, as we discuss later in this chapter.

Several generalizations can be made regarding spontaneous mutation rates. The **mutation rate** is defined as the likelihood that a gene will undergo a mutation in a single generation or in forming a single gamete. First, the rate of spontaneous mutation is exceedingly low for all organisms. Second, the rate varies between different organisms. Third, even within the same species, the spontaneous mutation rate varies from gene to gene.

Viral and bacterial genes undergo spontaneous mutation at an average of about 1 in 100 million (10^{-8}) replications or cell divisions (**Table 15.2**). Maize and *Drosophila* demonstrate rates several orders of magnitude higher. The genes studied in these groups average between 1 in 1,000,000 (10^{-6}) and 1 in 100,000 (10^{-5}) mutations per gamete formed. Some mouse genes are another order of magnitude higher in their spontaneous mutation rate: 1 in 100,000 to 1 in 10,000 (10^{-5} to 10^{-4}). It is not clear why such large variations occur in mutation rates.

The variation in rates between organisms may, in part, reflect the relative efficiencies of their DNA proofreading and repair systems. We will discuss these systems later in the chapter. Variation between genes in a given organism may be due to inherent differences in mutability in different regions of the genome. Some DNA sequences appear to be highly susceptible to mutation and are known as **mutation hot spots**.

In contrast to spontaneous mutations, mutations that result from the influence of exogenous factors are considered to be **induced mutations.** Induced mutations may be the result of either natural or artificial agents. For example, radiation from cosmic and mineral sources and ultraviolet (UV) radiation from the sun are energy sources to which most organisms are exposed and, as such, may be factors that cause induced mutations.

The earliest demonstration of the artificial induction of mutations occurred in 1927, when Hermann J. Muller reported that X rays could cause mutations in *Drosophila*. In 1928, Lewis J. Stadler reported that X rays had the same effect on DNA in barley. In addition to various forms of radiation, numerous natural and synthetic chemical agents are also mutagenic.

Spontaneous Germ-Line Mutation Rates in Humans

Until recently, scientists have studied human germ-line mutation rates by examining individual genes or small regions of the genome, particularly those genomic regions which, when mutated, contribute to specific genetic diseases. Now that whole-genome sequencing is becoming both rapid and economical, it is possible to examine entire genomes, both coding and noncoding regions, and to compare genomes from parents and offspring. These methods provide a more accurate estimate of mutation rates across the entire genome.

In 2012, a research group in Iceland sequenced the genomes of 78 parent/offspring sets, comprising 219 individuals, and compared the **single-nucleotide polymorphisms (SNPs)** (see Chapter 5) throughout their genomes.[1] Their data revealed that a newborn baby's

[1] Kong, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471–475.

genome contains an average of 60 new mutations, compared with those of his or her parents. Their research also revealed that the number of new mutations depends significantly on the age of the father at the time of conception. For example, when the father is 20 years old, he contributes approximately 25 new mutations to the child. When he is 40 years old, he contributes approximately 65 new mutations. In contrast, the mother contributes about 15 new mutations, at any age. The researchers estimated that the father contributes approximately 2 mutations per year of his age, with the mutation rate doubling every 16.5 years. The large proportion of mutations contributed by fathers is likely due to the fact that male germ cells go through more cell divisions during a lifetime than do female germ cells.

Of the 4933 new SNP mutations that were identified in this study, only 73 occurred within gene exons. Other studies have suggested that about 10 percent of single-nucleotide mutations lead to negative phenotypic changes. If so, then an average spontaneous mutation rate of 60 new mutations might yield about six deleterious phenotypic effects per generation.

Spontaneous Somatic Mutation Rates in Humans

During development, human cells undergo billions of cell divisions. In addition, some cell types, such as epithelial cells of the gut lining, continue to divide throughout life. At each division, these cells can acquire new mutations.

It is estimated that somatic cell mutation rates are between 4 and 25 times higher than those in germ-line cells. It is well accepted that somatic mutations are responsible for the development of most cancers. Cancer cells exhibit a wide range of types and numbers of somatic mutations—from a few to dozens of single nucleotide substitutions, as well as large chromosomal rearrangements. In addition, each type of cancer appears to exhibit characteristic mutations in genes specifically related to that cancer. These mutation patterns often reflect the actions of the cancer's causative agents such as specific mutagens or defects in DNA repair. We will discuss the effects of somatic mutations on the development of cancer in more detail later (see Chapter 24).

It is now known that somatic mutations can be responsible for other diseases besides cancer and can even become a source of new germ-line mutations. If a somatic mutation occurs in one cell very early in development, when the zygote contains only a few cells, that mutation may ultimately contribute to a large portion of the adult organism—a condition known as **somatic mosaicism** (see Chapter 7). The new mutation may create a new phenotype in the adult organism or may, if present in gonadal tissues, become a germ-line mutation that can be passed on to the next generation. Somatic mutations occurring early in development are estimated to lead to as many as 6 to 20 percent of Mendelian disorder cases.

The Fluctuation Test: Are Mutations Random or Adaptive?

One of the basic concepts in genetics is that mutations occur randomly and are not directed by the environment in which the organism finds itself—that is, mutations occur randomly rather than being the consequence of selective pressure. This concept has been verified by many experiments, including a classic experiment devised by Salvador Luria and Max Delbrück.

In 1943, Luria and Delbrück presented the first direct evidence that mutations do not occur as part of an adaptive mechanism, but instead take place spontaneously and randomly. Their experiment, known as the **Luria–Delbrück fluctuation test,** is an example of exquisite analytical and theoretical work.

Luria and Delbrück carried out their experiments using the *Escherichia coli*-T1 system. The T1 bacteriophage is a bacterial virus that infects *E. coli* cells; infected bacteria are lysed and die. Mutations can occur in *E. coli* cells that make the cells resistant to T1. To begin their experiment, Luria and Delbrück inoculated a large flask and a number of individual culture tubes with a few phage-sensitive *E. coli* cells and incubated all the cultures for several generations in the absence of any T1 bacteriophage. After growth, each small tube contained about 20 million cells, and the large flask was allowed to grow to a higher density. At this point, the bacteria in the smaller tubes were spread onto the surface of growth media containing T1 bacteriophage, in individual petri dishes. Similarly, several portions of the large flask (each portion also containing about 20 million cells) were spread onto T1-containing media. Any individual cells of *E. coli* that were resistant to the T1 bacteriophage grew and formed colonies on the petri plates. All other *E. coli* cells that were not resistant would die. The colonies present on the plates were then counted.

The experimental rationale for distinguishing between the two hypotheses (adaptive versus random mutation) was as follows.

Hypothesis 1: Adaptive Mutations If mutations occur adaptively (i.e., in response to the presence of T1 bacteriophage in the growth medium), every *E. coli* cell that is grown on a T1-containing growth medium would have a constant probability of acquiring a mutation that would give it resistance to T1. In this case, if a constant number of cells and T1 were present on each petri plate, a fairly constant number of resistant colonies should be observed from plate to plate and from experiment to experiment.

Hypothesis 2: Random Mutations If mutations occur randomly, mutations leading to resistance would occur, even in the absence of T1 bacteriophage, at a low fixed rate at any time during the incubation of each liquid culture. If a mutation occurred early in the incubation process, the subsequent growth of the mutant bacteria would produce a large number of resistant cells in the liquid culture. If a mutation occurred later in the incubation process, there would be fewer resistant cells. The random mutation hypothesis predicts that the number of resistant cells would fluctuate from experiment to experiment, and from small tube to small tube, reflecting the varying times at which the resistance mutations occurred in liquid culture. In contrast, each portion of the large culture flask (containing a stirred and homogeneous mixture of resistant and susceptible cells, thus serving as a control) would produce a constant number of resistant colonies from plate to plate.

The Results Luria and Delbrück's results revealed that there were a constant number of resistant cells in the large flask that had been mixed prior to plating on T1-containing medium. In contrast, there was a great deal of fluctuation in the numbers of resistant cells between independently incubated cultures. These data support the hypothesis that mutations arise randomly, even in the absence of selective pressure, and are inherited in a stable fashion.

Although the concept of spontaneous mutation has been accepted for some time, the possibility that organisms might also be capable of inducing specific mutations as a result of environmental pressures has long intrigued geneticists. Some recent and controversial research has suggested that under some stressful nutritional conditions such as starvation, bacteria may be capable of activating mechanisms that create a hypermutable state in genes that would, when mutated, enhance survival. The conclusions from these studies are still a source of debate, but they keep alive the interest in the possibility of adaptive mutation.

15.3 Spontaneous Mutations Arise from Replication Errors and Base Modifications

In this section, we will outline some of the processes that lead to spontaneous mutations. It is useful to keep in mind, however, that many of the DNA changes that occur during spontaneous mutagenesis also occur, at a higher rate, during induced mutagenesis.

DNA Replication Errors and Slippage

As we learned earlier (see Chapter 11), the process of DNA replication is imperfect. Occasionally, DNA polymerases insert incorrect nucleotides during replication of a strand of DNA. Although DNA polymerases can correct most of these replication errors using their inherent 3′ to 5′ exonuclease proofreading capacity, misincorporated nucleotides may persist after replication. If these errors are not detected and corrected by DNA repair mechanisms, they may lead to mutations. Replication errors due to mispairing predominantly lead to point mutations. The fact that bases can take several forms, known as **tautomers,** increases the chance of mispairing during DNA replication, as we will explain shortly.

In addition to mispairing and point mutations, DNA replication can lead to the introduction of small insertions or deletions. These mutations can occur when one strand of the DNA template loops out and becomes displaced during replication, or when DNA polymerase slips or stutters during replication—events termed **replication slippage.** If a loop occurs in the template strand during replication, DNA polymerase may miss the looped-out nucleotides, and a small deletion in the new strand will be introduced. If DNA polymerase repeatedly introduces nucleotides that are not present in the template strand, an insertion of one or more nucleotides will occur, creating an unpaired loop on the newly synthesized strand. Insertions and deletions may lead to frameshift mutations or amino acid insertions or deletions in the gene product.

Replication slippage can occur anywhere in the DNA but seems distinctly more common in regions containing tandemly repeated sequences. Repeat sequences are hot spots for DNA mutation and in some cases contribute to hereditary diseases, such as fragile-X syndrome and Huntington disease. The hypermutability of repeat sequences in noncoding regions of the genome is the basis for current methods of forensic DNA analysis.

In eukaryotes, at least four specialized DNA polymerases, known as translesion DNA polymerases, replicate DNA in regions of the genome that contain DNA damage. They are able to bypass the damaged nucleotides and continue replication but may introduce incorrect nucleotides and hence lead to mutations. (Translesion DNA polymerases are also discussed in Chapter 11.)

Tautomeric Shifts

Purines and pyrimidines can exist in tautomeric forms—that is, in alternate chemical forms that differ by the shift of a single proton in the molecule. The biologically important tautomers are the keto–enol forms of thymine and guanine and the amino–imino forms of cytosine and

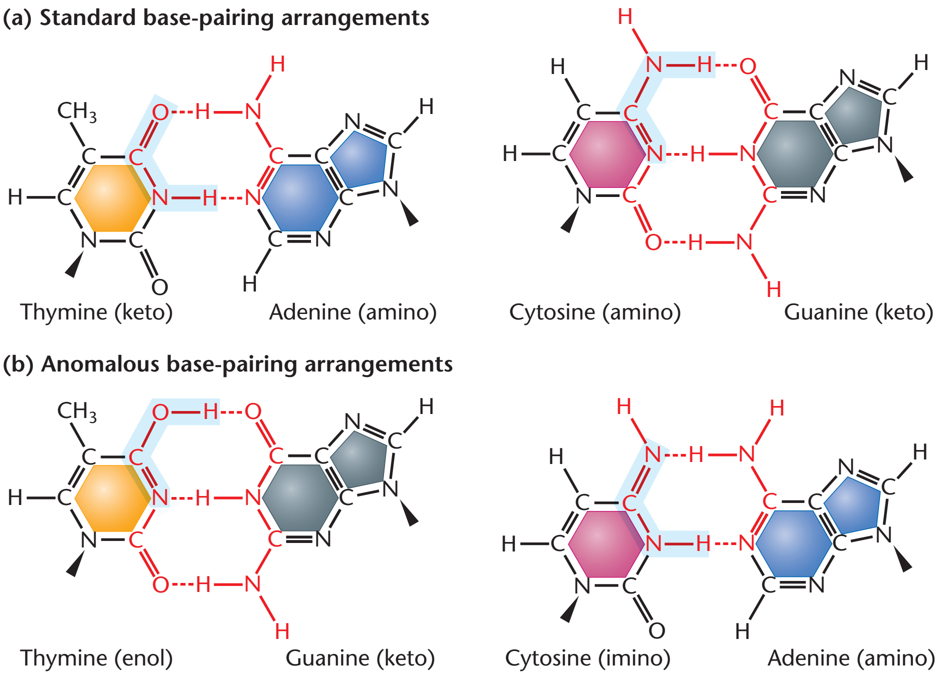


FIGURE 15.2 Examples of standard base-pairing relationships (a) compared with examples of the anomalous base pairing that occurs as a result of tautomeric shifts (b). The long triangles indicate the point at which each base bonds to a backbone sugar.

adenine. **Tautomeric shifts** change the covalent structure of the molecule, allowing hydrogen bonding with noncomplementary bases, and hence, may lead to permanent base-pair changes and mutations. Figure 15.2 compares normal base-pairing relationships with rare unorthodox pairings. Anomalous T-G and C-A pairs, among others, may be formed.

A mutation occurs during DNA replication when a transiently formed tautomer in the template strand pairs with a noncomplementary base. In the next round of replication, the "mismatched" members of the base pair are separated, and each becomes the template for its normal complementary base. The end result is a point mutation (Figure 15.3).

Depurination and Deamination

Some of the most common causes of spontaneous mutations are two forms of DNA base damage: depurination and deamination. **Depurination** is the loss of one of the nitrogenous bases in an intact double-helical DNA molecule. Most frequently, the base is either guanine or adenine—in other words, a purine. These bases may be lost if the glycosidic bond linking the 1′-C of the deoxyribose and the number 9 position of the purine ring is broken, leaving an **apurinic site** on one strand of the DNA. Geneticists estimate that thousands of such spontaneous lesions are formed daily in the DNA of mammalian cells in culture. If apurinic sites are not repaired, there will be no base at that position to act as a template during DNA replication. As a result, DNA polymerase may introduce a nucleotide at random at that site.

In **deamination,** an amino group in cytosine or adenine is converted to a keto group. In these cases, cytosine is converted to uracil, and adenine is changed to the guanine-resembling compound hypoxanthine (Figure 15.4). The major effect of these changes is an alteration in the base-pairing specificities of these two bases during DNA replication. For example, cytosine normally pairs with guanine. Following its conversion to uracil, which pairs with adenine, the original G-C pair is converted to an A-U pair and then, in the next replication, is converted to an A-T pair. When adenine is deaminated, the original A-T pair is ultimately converted to a G-C pair because hypoxanthine pairs naturally with cytosine, which then pairs with guanine in the next replication. Deamination may occur spontaneously or as a result of treatment with chemical mutagens such as nitrous acid (HNO_2).

Oxidative Damage

DNA may also suffer damage from the by-products of normal cellular processes. These by-products include reactive oxygen species (electrophilic oxidants) that are generated during normal aerobic respiration. For example, superoxides (O_2^-) hydroxyl radicals ($^{\bullet}OH$), and hydrogen peroxide (H_2O_2) are created during cellular metabolism and are constant threats to the integrity of DNA. Such **reactive oxidants,** also generated by exposure to high-energy radiation, can produce more than 100 different types of chemical modifications in DNA, including modifications to bases, loss of bases, and single-stranded breaks.

Transposable Elements

Transposable elements are DNA sequences that can move within genomes. These mobile elements are present in the genomes of all organisms, from bacteria to humans, and often constitute large portions of these genomes. Transposable elements can act as naturally occurring mutagens. If in moving to a new location they insert themselves into the coding region of a gene, they can alter the reading frame or introduce stop codons. If they insert into the regulatory region of a gene, they can disrupt proper expression of the gene. Transposable elements can also create chromosomal damage, including double-stranded breaks, inversions, and translocations. Transposable elements are described in detail later in this chapter (Section 15.8).

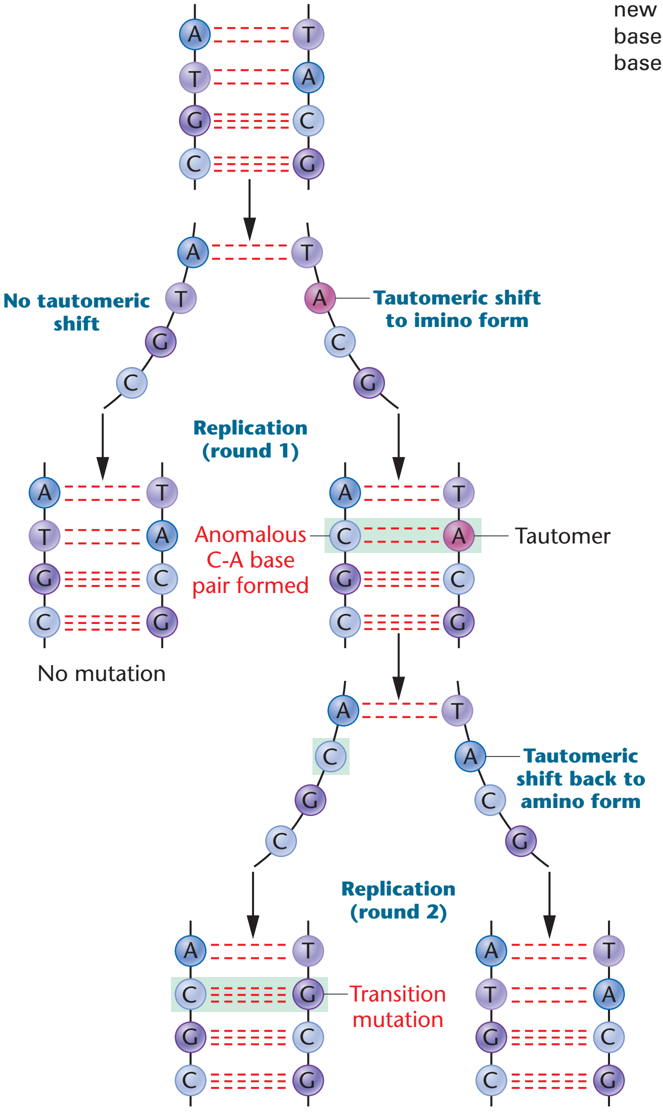


FIGURE 15.3 Formation of an A-T to G-C transition mutation as a result of a transient tautomeric shift in adenine.

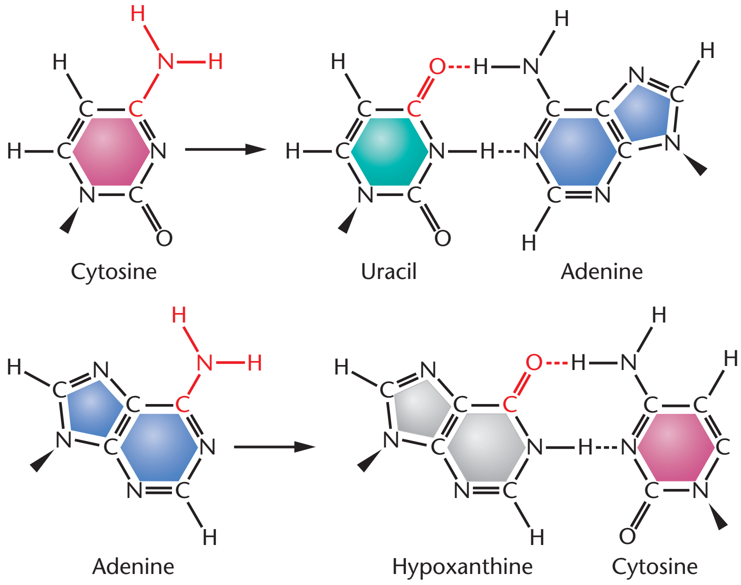


FIGURE 15.4 Deamination of cytosine and adenine, leading to new base pairing and mutation. Cytosine is converted to uracil, which base-pairs with adenine. Adenine is converted to hypoxanthine, which base-pairs with cytosine.

NOW SOLVE THIS

15.2 One of the most famous cases of an X-linked recessive mutation in humans is that of hemophilia found in the descendants of Britain's Queen Victoria. The pedigree of the royal family indicates that Victoria was heterozygous for the trait; however, her father was not affected, and no other member of her maternal line appeared to carry the mutation. What are some possible explanations of how the mutation arose? What types of mutations could lead to the disease?

- **Hint:** *This problem asks you to determine the sources of new mutations. The key to its solution is to consider the ways in which mutations occur, the types of cells in which they can occur, and how they are inherited.*

15.4 Induced Mutations Arise from DNA Damage Caused by Chemicals and Radiation

All cells on Earth are exposed to a plethora of agents called **mutagens,** which have the potential to damage DNA and cause induced mutations. Some of these agents, such as some fungal toxins, cosmic rays, and UV light, are natural components of our environment. Others, including some industrial pollutants, medical X rays, and chemicals within tobacco smoke, can be considered as unnatural or human-made additions to our modern world. On the positive

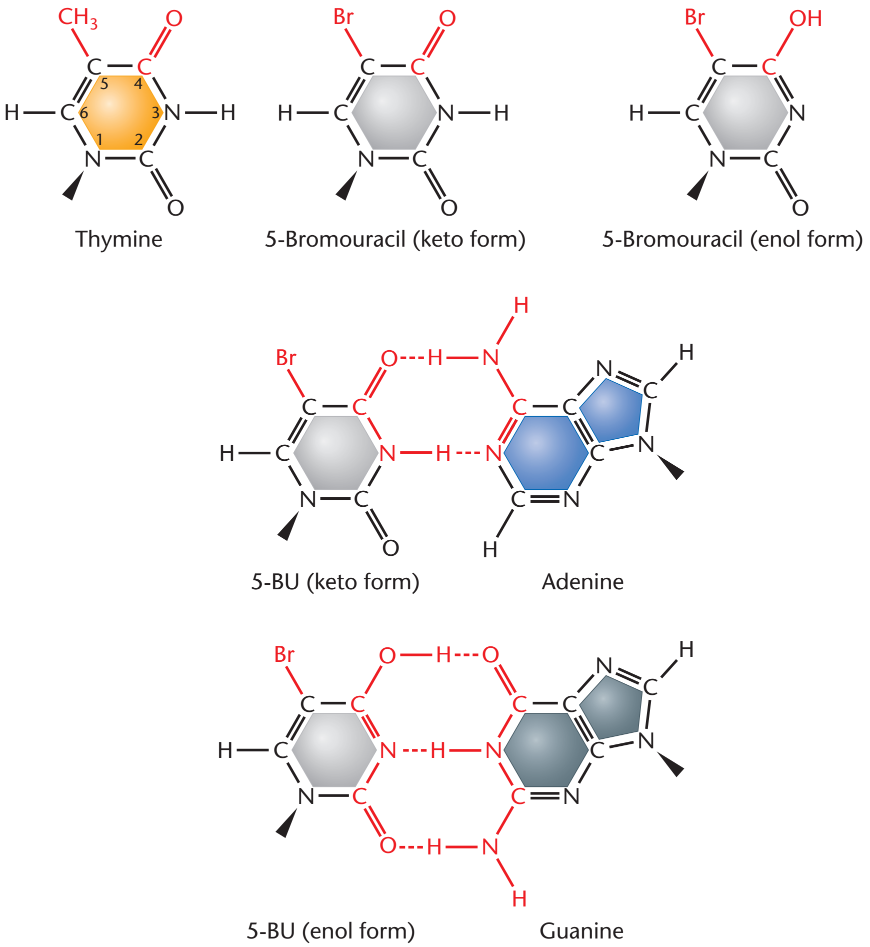


FIGURE 15.5 Similarity of the chemical structure of 5-bromouracil (5-BU) and thymine. In the common keto form, 5-BU base-pairs normally with adenine, behaving as a thymine analog. In the rare enol form, it pairs anomalously with guanine.

side, geneticists have harnessed some mutagens for use in analyzing genes and gene functions. The mechanisms by which some of these natural and unnatural agents lead to mutations are outlined in this section.

Base Analogs

One category of mutagenic chemicals is **base analogs,** compounds that can substitute for purines or pyrimidines during nucleic acid biosynthesis. For example, **5-bromouracil (5-BU),** a derivative of uracil, behaves as a thymine analog but with a bromine atom substituted at the number 5 position of the pyrimidine ring. (If 5-BU is chemically linked to deoxyribose, the nucleoside analog **bromodeoxyuridine (BrdU)** is formed.) **Figure 15.5** compares the structure of 5-BU with that of thymine. The presence of the bromine atom in place of the methyl group increases the probability that a tautomeric shift will occur. If BrdU is incorporated into DNA in place of thymidine and a tautomeric shift to the enol form of 5-BU occurs, 5-BU base-pairs with guanine. After one round of replication, an A-T to G-C transition results. Furthermore, the presence of 5-BU within DNA increases the sensitivity of the molecule to UV light, which itself is mutagenic.

There are other base analogs that are mutagenic. For example, **2-amino purine (2-AP)** can act as an analog of adenine. In addition to its base-pairing affinity with thymine, 2-AP can also base-pair with cytosine, leading to possible transitions from A-T to G-C following replication.

Alkylating, Intercalating, and Adduct-Forming Agents

A number of naturally occurring and human-made chemicals alter the structure of DNA and cause mutations. The sulfur-containing mustard gases, discovered during World War I, were some of the first chemical mutagens identified in chemical warfare studies. Mustard gases are **alkylating agents**—that is, they donate an alkyl group, such as CH_3 or CH_2CH_3, to amino or keto groups in nucleotides. Ethylmethane sulfonate (EMS), for example, alkylates the keto groups in the number 6 position of guanine and in the number 4 position of thymine. As with base analogs, base-pairing affinities are altered, and transition mutations result. For example, 6-ethylguanine acts as an analog of adenine and pairs with thymine (**Figure 15.6**).

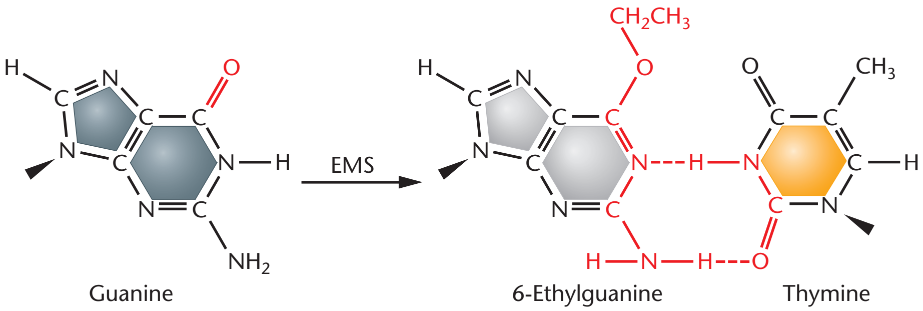


FIGURE 15.6 Conversion of guanine to 6-ethylguanine by the alkylating agent ethylmethane sulfonate (EMS). The 6-ethylguanine base-pairs with thymine.

Intercalating agents are chemicals that have dimensions and shapes that allow them to wedge between the base pairs of DNA. Wedged intercalating agents cause base pairs to distort and DNA strands to unwind. These changes in DNA structure affect many functions including transcription, replication, and repair. Deletions and insertions occur during DNA replication and repair, leading to frame-shift mutations.

Some intercalating agents are used as DNA stains. An example is ethidium bromide, a fluorescent compound that is commonly used in molecular biology laboratories to visualize DNA during purifications and gel electrophoresis. The mutagenic characteristics of both ethidium bromide and the UV light used to visualize its fluorescence mean that this chemical must be used with caution. Other intercalating agents are used for cancer chemotherapy. Examples are doxorubicin, which is used to treat Hodgkin lymphoma, and dactinomycin, which is used to treat a variety of sarcomas. Because cancer cells undergo DNA replication more frequently than noncancer cells, they are more sensitive than normal cells to the killing effects of these chemotherapeutic agents.

Another group of chemicals that cause mutations are known as **adduct-forming agents.** A DNA adduct is a substance that covalently binds to DNA, altering its conformation and interfering with replication and repair. Two examples of adduct-forming substances are acetaldehyde (a component of cigarette smoke) and heterocyclic amines (HCAs). HCAs are cancer-causing chemicals that are created during the cooking of meats such as beef, chicken, and fish. HCAs are formed at high temperatures from amino acids and creatine. Many HCAs covalently bind to guanine bases. At least 17 different HCAs have been linked to the development of cancers, such as those of the stomach, colon, and breast.

Ultraviolet Light

All electromagnetic radiation consists of energetic waves that we define by their different wavelengths (**Figure 15.7**). The full range of wavelengths is referred to as the **electromagnetic spectrum**, and the energy of any radiation in the spectrum varies inversely with its wavelength. Waves in the range of visible light and longer are benign when they interact with most organic molecules. However, waves of shorter length than visible light, being inherently more energetic, have the potential to disrupt organic molecules.

As we know (see Chapter 10), purines and pyrimidines absorb **ultraviolet (UV) radiation** most intensely at a wavelength of about 260 nm. Although Earth's ozone layer absorbs the most dangerous types of UV radiation, sufficient UV radiation can induce thousands of DNA lesions per hour in any cell exposed to this radiation. One major effect of UV radiation on DNA is the creation of **pyrimidine dimers**—chemical species consisting of two identical pyrimidines—particularly ones consisting of two thymidine residues (**Figure 15.8**). The dimers distort the DNA conformation and inhibit normal replication. As a result, errors can be introduced in the base sequence of DNA during replication through the actions of error-prone DNA polymerases. When UV-induced dimerization is extensive, it is responsible (at least in part) for the killing effects of UV radiation on cells.

Ionizing Radiation

As noted above, the energy of radiation varies inversely with wavelength. Therefore, **X rays, gamma rays,** and **cosmic rays** are more energetic than UV radiation (Figure 15.7).

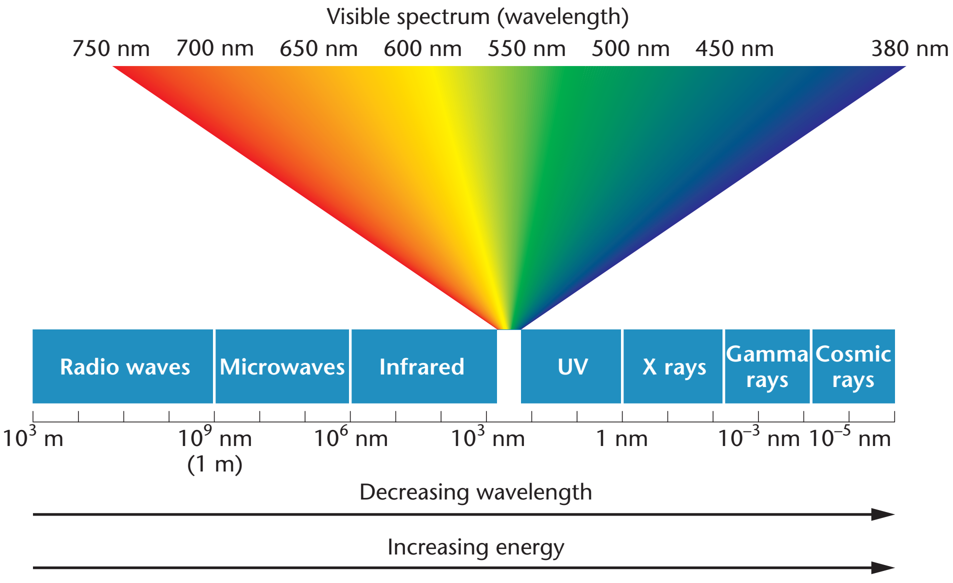


FIGURE 15.7 The regions of the electromagnetic spectrum and their associated wavelengths.

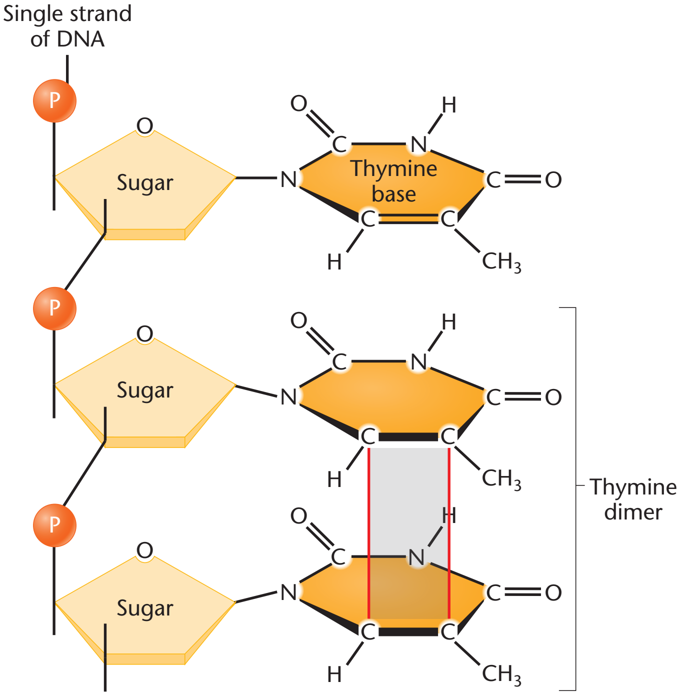


FIGURE 15.8 Depiction of a thymine dimer induced by UV radiation. The covalent crosslinks (shown in red) occur between carbon atoms of the pyrimidine rings.

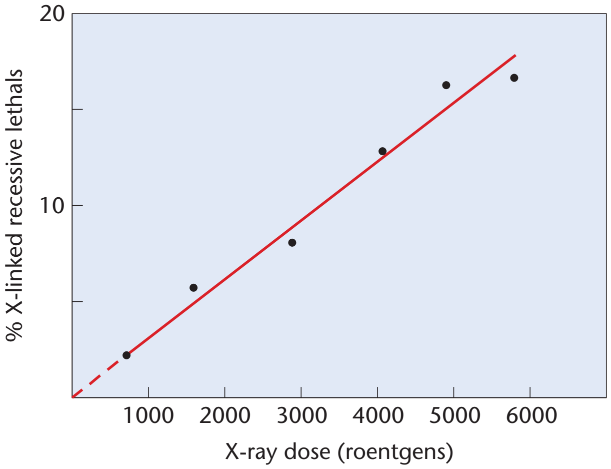


FIGURE 15.9 Plot of the percentage of X-linked recessive mutations induced in *Drosophila* by increasing doses of X rays. If extrapolated, the graph intersects the zero axis as shown by the dashed line.

As a result, they penetrate deeply into tissues, causing ionization of the molecules encountered along the way. Hence, this type of radiation is called **ionizing radiation**.

As ionizing radiation penetrates cells, stable molecules and atoms are transformed into **free radicals**—chemical species containing one or more unpaired electrons. Free radicals can directly or indirectly affect the genetic material, altering purines and pyrimidines in DNA, breaking phosphodiester bonds, disrupting the integrity of chromosomes, and producing a variety of chromosomal aberrations, such as deletions, translocations, and chromosomal fragmentation.

Given the capacity of ionizing radiation to cause serious genetic damage, it is important to consider what levels of radiation are mutagenic in humans and what sources of ionizing radiation cause the most damage in everyday life. **Figure 15.9** shows a graph of the percentage of induced X-linked recessive lethal mutations versus the dose of X rays administered in *Drosophila*. There is a linear relationship between X-ray dose and the induction of mutation; for each doubling of the dose, twice as many mutations are induced. Because the line intersects near the zero axis, this graph suggests that even very small doses of radiation are mutagenic.

In humans, one way to assess the potential biological consequences of radiation-induced DNA damage is to examine the effects of ionizing radiation on rates of leukemia. It is well known that ionizing radiation increases the risk of developing cancers such as leukemias. For example, the survivors of the atomic bombings of Nagasaki and Hiroshima in 1945 had an increase in the incidence of leukemias, with higher radiation doses resulting in higher cancer rates. Although the effects of such high-dose exposures to ionizing radiation are clear, it has been difficult to show the effects of lower doses. However, recent studies of long-term exposures to low levels of radiation have confirmed that even very low doses can have detectable, although small, effects on leukemia rates. These studies show that exposure to an additional 10 mSv (millisieverts; a measure of radiation effect on tissues) of ionizing radiation can raise a person's risk of leukemia by 0.002 percent. To provide context, normal background radiation is approximately 2–3 mSv per year from all sources.

Although it is often assumed that radiation from artificial sources such as nuclear power plant waste and medical X rays are the most significant sources of radiation exposure for humans, scientific data indicate otherwise. Scientists estimate that less than 20 percent of human radiation exposure arises from human-made sources. **Figure 15.10** summarizes the annual radiation exposure for humans residing in the United States. As these data indicate, the greatest radiation exposure comes from radon gas, cosmic rays, and natural soil radioactivity. More than half of human-made radiation exposure comes from medical X rays and radioactive pharmaceuticals.

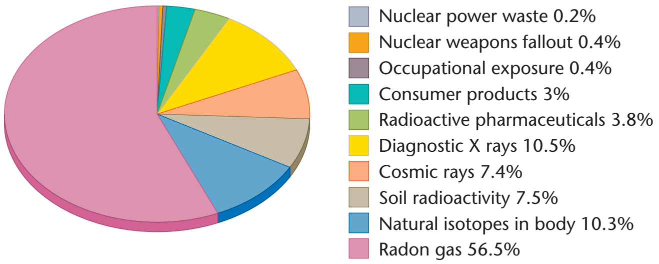


FIGURE 15.10 Chart showing average yearly dose of radiation from natural and human-made sources.

NOW SOLVE THIS

15.3 The cancer drug melphalan is an alkylating agent of the mustard gas family. It acts in two ways: by causing alkylation of guanine bases and by cross linking DNA strands together. Describe two ways in which melphalan might kill cancer cells. What are two ways in which cancer cells could repair the DNA-damaging effects of melphalan?

- **Hint:** *This problem asks you to consider the effect of the alkylation of guanine on base pairing during DNA replication. The key to its solution is to consider the effects of mutations on cellular processes that allow cells to grow and divide. In Section 15.6, you will learn about the ways in which cells repair the types of mutations introduced by alkylating agents.*

15.5 Single-Gene Mutations Cause a Wide Range of Human Diseases

Although most human genetic diseases are **polygenic**—that is, caused by variations in several genes—even a single base-pair change in one of the approximately 20,000 human genes can lead to a serious inherited disorder. These **monogenic** diseases can be caused by many different types of single-gene mutations. **Table 15.3** lists some examples of the types of single-gene mutations that can lead to serious genetic diseases. A comprehensive database of human genes, mutations, and disorders is available in the Online Mendelian Inheritance in Man (OMIM) database (described in the Exploring Genomics feature in Chapter 3). As of 2017, the OMIM database has cataloged approximately 6000 human phenotypes for which the molecular basis is known.

Geneticists estimate that approximately 30 percent of mutations that cause human diseases are single base-pair changes that create nonsense mutations. These mutations not only code for a prematurely terminated protein product, but also trigger rapid decay of the mRNA. Many more mutations are missense mutations that alter the amino acid sequence of a protein and frameshift mutations that alter the protein sequence and create internal nonsense codons. Other common disease-associated mutations affect the sequences of gene promoters, mRNA splicing signals, and other noncoding sequences that affect transcription, processing, and stability of mRNA or protein. One recent study showed that about 15 percent of all point mutations that cause human genetic diseases result in abnormal mRNA splicing. Approximately 85 percent of these splicing mutations alter the sequence of 5′ and 3′ splice signals. The remainder create new splice sites within the gene. Splicing defects often result in degradation of the abnormal mRNA or creation of abnormal protein products.

Single-Gene Mutations and β-Thalassemia

Although some single-gene diseases, such as sickle-cell anemia (introduced in Chapter 14), are caused by one specific base-pair change within a gene, most are caused by any of a large number of different mutations. The mutation profile associated with β-thalassemia provides an example of the latter, more common, type of monogenic disease.

β-thalassemia is an inherited autosomal recessive blood disorder resulting from a reduction or absence of hemoglobin. It is the most common single-gene disease in the world, affecting people worldwide, but especially populations in Mediterranean, North African, Middle Eastern, Central Asian, and Southeast Asian countries.

People with β-thalassemia have varying degrees of anemia—from severe to mild—with symptoms including weakness, delayed development, jaundice, enlarged organs, and often a need for frequent blood transfusions.

TABLE 15.3 Examples of Human Disorders Caused by Single-Gene Mutations

Type of Mutation	Disorder	Molecular Change
Missense	Achondroplasia	Glycine to arginine at position 380 of *FGFR3* gene
Nonsense	Marfan syndrome	Tyrosine to STOP codon at position 2113 of *fibrillin-1* gene
Insertion	Familial hypercholesterolemia	Various short insertions throughout the *LDLR* gene
Deletion	Cystic fibrosis	Three-base-pair deletion of phenylalanine codon at position 508 of *CFTR* gene
Trinucleotide repeat expansions	Huntington disease	>40 repeats of (CAG) sequence in coding region of *Huntingtin* gene

Mutations in the β-*globin* gene (*HBB* gene) cause β-thalassemia. The *HBB* gene encodes the 146-amino-acid β-globin polypeptide. Two β-globin polypeptides associate with two α-globin polypeptides to form the adult hemoglobin tetramer. The *HBB* gene spans 1.6 kilobases of DNA on the short arm of chromosome 11. It is made up of three exons and two introns.

Scientists have discovered approximately 400 different mutations in the *HBB* gene that cause β-thalassemia, although most cases worldwide are associated with only about 20 of these mutations. Most mutations change a single nucleotide within or surrounding the *HBB* gene or create small insertions and deletions. In addition, each population affected by β-thalassemia has a unique mix of mutations. For example, the most prevalent mutation in a Sardinian population—a mutation that accounts for more than 95 percent of cases—results in a single base-pair change at codon 39, creating a nonsense mutation and premature termination of the β-globin polypeptide. In contrast, a study of β-thalassemia mutations in a population from the former Yugoslavia revealed 14 different mutations, with only three (all in intron 1 splice signals) accounting for 75 percent of cases.

The types of mutations that cause β-thalassemia not only affect the β-globin amino acid sequence (missense, nonsense, and frameshift mutations), but also alter *HBB* transcription efficiency, mRNA splicing and stability, translation, and protein stability.

Table 15.4 provides a summary of the types of single-gene mutations that cause β-thalassemia. More than half of these mutations are single base-pair changes, and the remainder are short insertions, deletions, and duplications.

Mutations Caused by Expandable DNA Repeats

Beginning in about 1990, molecular analyses of the genes responsible for a number of inherited human disorders revealed a remarkable set of observations. Researchers discovered that some mutant genes contain an expansion of **trinucleotide repeat sequences**—specific short DNA sequences repeated many times. Normal individuals have a low number of repetitions of these sequences; however, individuals with over 20 different human disorders appear to have abnormally large numbers of repeat sequences—in some cases, over 200—within and surrounding specific genes.

Examples of diseases associated with these trinucleotide repeat expansions are fragile-X syndrome (discussed in detail in Chapter 8), myotonic dystrophy, and Huntington disease (discussed in Chapter 4). When trinucleotide repeats such as $(CAG)_n$ occur within a coding region, they can be translated into long tracks of glutamine. These glutamine tracks may cause the proteins to aggregate abnormally. When the repeats occur outside coding regions, but within the mRNA, it is thought that the mRNAs may act as "toxic" RNAs that bind to important regulatory proteins, sequestering them away from their normal functions in the cell. Another possible consequence of long trinucleotide repeats is that the regions of DNA containing the repeats may become abnormally methylated, leading to silencing of gene transcription.

The mechanisms by which the repeated sequences expand from generation to generation are of great interest. It is thought that expansion may result from either errors during DNA replication or errors during DNA damage repair. Whatever the cause may be, the presence of these short and unstable repeat sequences seems to be prevalent in humans and in many other organisms.

TABLE 15.4 Types of Mutations in the *HBB* Gene That Cause β-Thalassemia

Gene Region Affected	Number of Mutations Known	Description
5′ upstream region	22	Single base-pair mutations occur between −101 and −25 upstream from transcription start site. For example, a T → A transition in the TATA sequence at −30 results in decreased gene transcription and severe disease.
mRNA CAP site	1	Single base-pair mutation (A → C transversion) at +1 position leads to decreased levels of mRNA.
5′ untranslated region	3	Single base-pair mutations at +20, +22, and +33 cause decreases in transcription and translation and mild disease.
ATG translation initiation codon	7	Single base-pair mutations alter the mRNA AUG sequence, resulting in no translation and severe disease.
Exons 1, 2, and 3 coding regions	36	Single base-pair missense and nonsense mutations, and mutations that create abnormal mRNA splice sites. Disease severity varies from mild to extreme.
Introns 1 and 2	38	Single base-pair transitions and transversions that reduce or abolish mRNA splicing and create abnormal splice sites that affect mRNA stability. Most cause severe disease.
Polyadenylation site	6	Single base-pair changes in the AATAAA sequence reduce the efficiency of mRNA cleavage and polyadenylation, yielding long mRNAs or unstable mRNAs. Disease is mild.
Throughout and surrounding the *HBB* gene	> 100	Short insertions, deletions, and duplications that alter coding sequences, create frameshift stop codons, and alter mRNA splicing.

15.6 Organisms Use DNA Repair Systems to Counteract Mutations

Living systems have evolved a variety of elaborate repair systems that counteract both spontaneous and induced DNA damage. These **DNA repair** systems are absolutely essential to the maintenance of the genetic integrity of organisms and, as such, to the survival of organisms on Earth. The balance between mutation and repair results in the observed mutation rates of individual genes and organisms. Of foremost interest in humans is the ability of these systems to counteract genetic damage that would otherwise result in genetic diseases and cancer. The link between defective DNA repair and cancer susceptibility is described later (see Chapter 24).

In 2015, two prestigious scientific awards were granted to scientists whose research has contributed to our understanding of the mechanisms of DNA repair. The Nobel Prize in Chemistry was awarded to Tomas Lindahl, Paul Modrich, and Aziz Sancar for their ground-breaking insights into the ways that cells detect and repair DNA damage—specifically the processes of base excision repair, mismatch repair, and nucleotide excision repair. The Albert Lasker Basic Medical Research Award honored the work of Evelyn Witkin and Stephen Elledge. Witkin's research uncovered the mechanisms and proteins that protect bacteria such as *E. coli* from radiation-induced DNA damage—a response mechanism known as SOS repair. Elledge's research provided insights into the proteins and biochemical pathways that are essential for DNA repair in eukaryotes.

We now embark on a review of these and other DNA repair mechanisms, with the emphasis on the major approaches that organisms use to counteract genetic damage.

Proofreading and Mismatch Repair

Some of the most common types of mutations arise during DNA replication when an incorrect nucleotide is inserted by DNA polymerase. The major DNA synthesizing enzyme in bacteria (**DNA polymerase III**) makes an error approximately once every 100,000 insertions, leading to an error rate of 10^{-5}. Fortunately, DNA polymerase proofreads each step, catching 99 percent of those errors. If an incorrect nucleotide is inserted during polymerization, the enzyme can recognize the error and "reverse" its direction. It then behaves as a 3′ to 5′ exonuclease, cutting out the incorrect nucleotide and replacing it with the correct one. This improves the efficiency of replication 100-fold, creating only 1 mismatch in every 10^7 insertions, for a final error rate of 10^{-7}.

To cope with errors such as base–base mismatches, small insertions, and deletions that remain after proofreading, another mechanism, called **mismatch repair (MMR),** may be activated. During MMR, the mismatches are detected, the incorrect nucleotide is removed, and the correct nucleotide is inserted in its place. But how does the repair system recognize which nucleotide is correct (on the template strand) and which nucleotide is incorrect (on the newly synthesized strand)? If a mismatch is recognized but no such discrimination occurs, the excision will be random, and the strand bearing the correct base will be clipped out 50 percent of the time. Hence, strand discrimination is a critical step.

The process of strand discrimination has been elucidated in some bacteria, including *E. coli*, and is based on **DNA methylation.** These bacteria contain an enzyme, **DNA adenine methylase,** which recognizes the DNA sequence

5′-GATC-3′
3′-CTAG-5′

as a substrate, adding a methyl group to each of the adenine residues during DNA replication.

Following replication, the newly synthesized DNA strand remains temporarily unmethylated, as the DNA adenine methylase lags behind the DNA polymerase. Prior to adenine methylation, the repair enzymes mentioned below are able to recognize any mismatch that is introduced on the newly synthesized (unmethylated) DNA strand and bind to the unmethylated strand. An **endonuclease** enzyme creates a nick in the backbone of the unmethylated DNA strand, either 5′ or 3′ to the mismatch. An **exonuclease** unwinds and degrades the nicked DNA strand, until the region of the mismatch is reached. Finally, DNA polymerase fills in the gap created by the exonuclease, using the correct DNA strand as a template. DNA ligase then seals the gap.

A series of *E. coli* gene products, MutH, MutL, and MutS, as well as exonucleases, DNA polymerase III, and DNA ligase, are involved in MMR. Mutations in the *mutH*, *mutL*, and *mutS* genes result in bacterial strains deficient in MMR. While the preceding mechanism occurs in *E. coli*, similar mechanisms involving homologous proteins exist in yeast and in mammals.

In humans, mutations in genes that code for DNA MMR proteins (such as *hMSH2* and *hMLH1,* which are the human equivalents of the *mutS* and *mutL* genes of *E. coli*) are associated with the hereditary nonpolyposis colon cancer. MMR defects are commonly found in other cancers, such as leukemias, lymphomas, and tumors of the ovary, prostate, and endometrium. Cells from these cancers show genome-wide increases in the rate of spontaneous mutation. The link between defective MMR and cancer is supported by experiments with mice. Mice that are engineered to have deficiencies in MMR genes accumulate large numbers of mutations and are cancer-prone.

Postreplication Repair and the SOS Repair System

Another type of DNA repair system, called **postreplication repair,** responds *after* damaged DNA has escaped repair and has failed to be completely replicated. As illustrated in Figure 15.11, when DNA bearing a lesion of some sort (such as a pyrimidine dimer) is being replicated, DNA polymerase may stall at the lesion and then skip over it, leaving an unreplicated gap on the newly synthesized strand. To correct the gap, RecA protein directs a recombinational exchange with the corresponding region on the undamaged parental strand of the same polarity (the "donor" strand). When the undamaged segment of the donor strand DNA replaces the gapped segment, a gap is created on the donor strand. The gap can be filled by repair synthesis as replication proceeds. Because a recombinational event is involved in this type of DNA repair, it is considered to be a form of **homologous recombination repair**.

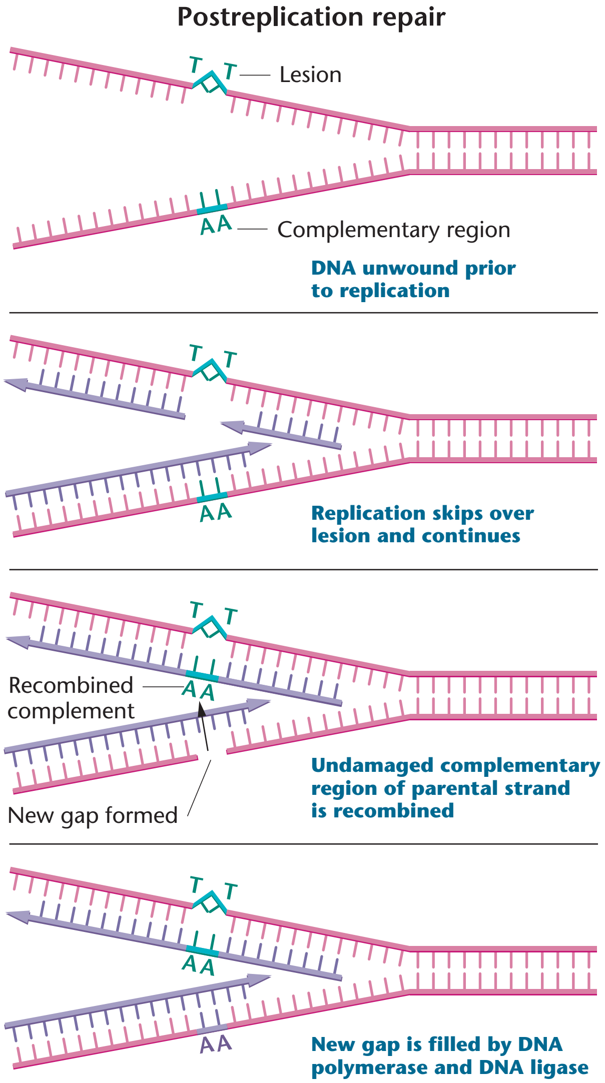

Postreplication repair
T T
Lesion
A A
Complementary region
DNA unwound prior to replication
T T
A A
Replication skips over lesion and continues
T T
Recombined complement
A A
New gap formed
Undamaged complementary region of parental strand is recombined
T T
A A
A A
New gap is filled by DNA polymerase and DNA ligase

FIGURE 15.11 Postreplication repair occurs if DNA replication has skipped over a lesion such as a thymine dimer. Through the process of recombination, the correct complementary sequence is recruited from the parental strand and inserted into the gap opposite the lesion. The new gap is filled by DNA polymerase and DNA ligase.

Another postreplication repair pathway, the *E. coli* **SOS repair system,** also responds to damaged DNA, but in a different way. In the presence of a large number of unrepaired DNA mismatches and gaps, the bacteria can induce expression of about 20 genes (including *lexA*, *recA*, and *uvr*) whose products allow DNA replication to occur even in the presence of DNA lesions. This type of repair is a last resort to minimize DNA damage, hence its name. During SOS repair, DNA synthesis becomes error-prone, inserting random and possibly incorrect nucleotides in places that would normally stall DNA replication. As a result, SOS repair itself becomes mutagenic—although it may allow the cell to survive DNA damage that would otherwise kill it.

Photoreactivation Repair: Reversal of UV Damage

As was illustrated in Figure 15.8, UV light introduces mutations by the creation of pyrimidine dimers. UV-induced damage to *E. coli* DNA can be partially reversed if, following irradiation, the cells are exposed briefly to visible light, especially in the blue range of the visible spectrum. The process is dependent on the activity of a protein called **photoreactivation enzyme (PRE)** or **photolyase.** The enzyme's mode of action is to cleave the cross-linking bonds between thymine dimers (Figure 15.12). Although the enzyme will associate with a thymine dimer in the dark, it must absorb a photon of blue light to cleave the dimer. In spite of its ability to reduce the number of UV-induced mutations, **photoreactivation repair** is not absolutely essential in *E. coli*; we know this because a mutation creating a null allele in the gene coding for PRE is not lethal. The enzyme is also detectable in many organisms, including other bacteria, fungi, plants, and some vertebrates—though not in humans. Humans and other organisms that lack photoreactivation repair must rely on other repair mechanisms to reverse the effects of UV radiation.

Base and Nucleotide Excision Repair

A number of light-independent DNA repair systems exist in all bacteria and eukaryotes. The basic mechanisms involved in these types of repair—collectively referred to as **excision repair** or cut-and-paste mechanisms—consist of the following three steps.

1. The damage, distortion, or error present on one of the two strands of the DNA helix is recognized and enzymatically clipped out by an endonuclease. Excisions in the phosphodiester backbone usually include

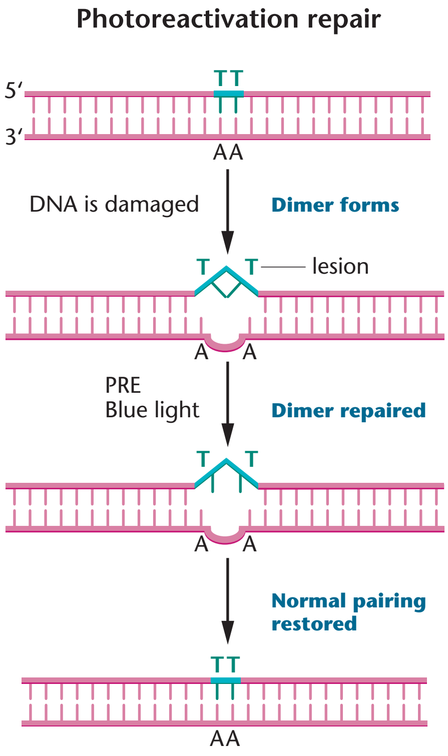


FIGURE 15.12 Damaged DNA repaired by photoreactivation repair. The bond creating the thymine dimer is cleaved by the photoreactivation enzyme (PRE), which must be activated by blue light in the visible spectrum.

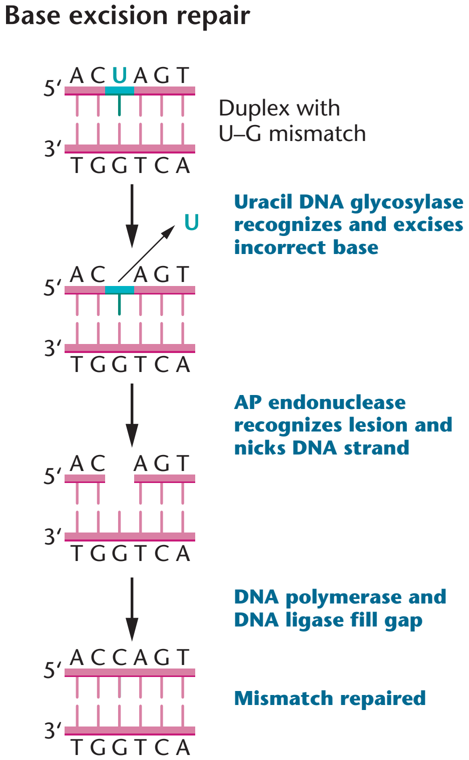


FIGURE 15.13 Base excision repair (BER) accomplished by uracil DNA glycosylase, AP endonuclease, DNA polymerase, and DNA ligase. Uracil is recognized as a noncomplementary base, excised, and replaced with the complementary base (C).

a number of nucleotides adjacent to the error as well, leaving a gap on one strand of the helix.

2. A DNA polymerase fills in the gap by inserting nucleotides complementary to those on the intact strand, which it uses as a replicative template. The enzyme adds these nucleotides to the free 3′-OH end of the clipped DNA. In *E. coli*, this step is usually performed by DNA polymerase I.

3. DNA ligase seals the final "nick" that remains at the 3′-OH end of the last nucleotide inserted, closing the gap.

There are two types of excision repair: base excision repair and nucleotide excision repair. **Base excision repair (BER)** corrects DNA that contains incorrect base pairings due to the presence of chemically modified bases or uridine nucleosides that are inappropriately incorporated into DNA or created by deamination of cytosine. The first step in the BER pathway involves the recognition of an inappropriately paired base by enzymes called **DNA glycosylases.** There are a number of DNA glycosylases, each of which recognizes a specific base. For example, the enzyme uracil DNA glycosylase recognizes the presence of uracil in DNA (Figure 15.13). DNA glycosylases first cut the glycosidic bond between the target base and its sugar, creating an **apyrimidinic** (or apurinic) **site**. The sugar with the missing base is then recognized by an enzyme called **AP endonuclease.** The AP endonuclease makes cuts in the phosphodiester backbone at the apyrimidinic or apurinic site. The gap is filled by DNA polymerase and DNA ligase.

Although much has been learned about the mechanisms of BER in *E. coli*, BER systems have also been detected in eukaryotes from yeast to humans. Experimental evidence shows that both mouse and human cells that are defective in BER activity are hypersensitive to the killing effects of gamma rays and oxidizing agents.

Nucleotide excision repair (NER) pathways repair "bulky" lesions in DNA that alter or distort the double helix. These lesions include the UV-induced pyrimidine dimers and DNA adducts discussed previously.

The NER pathway (Figure 15.14) was first discovered in 1964 by Paul Howard-Flanders and coworkers, who isolated several independent *E.coli* mutants that are sensitive to UV radiation. One group of genes was designated *uvr* (ultraviolet repair) and included the *uvrA, uvrB,* and *uvrC* mutations. In the NER pathway, the *uvr* gene products are involved in recognizing and clipping out lesions in the DNA. Usually, a specific number of nucleotides are clipped out around both sides of the lesion. In *E. coli,* usually a total of 13 nucleotides are removed, including the lesion. The repair is then completed by DNA polymerase I and DNA ligase, in a manner similar to that occurring in

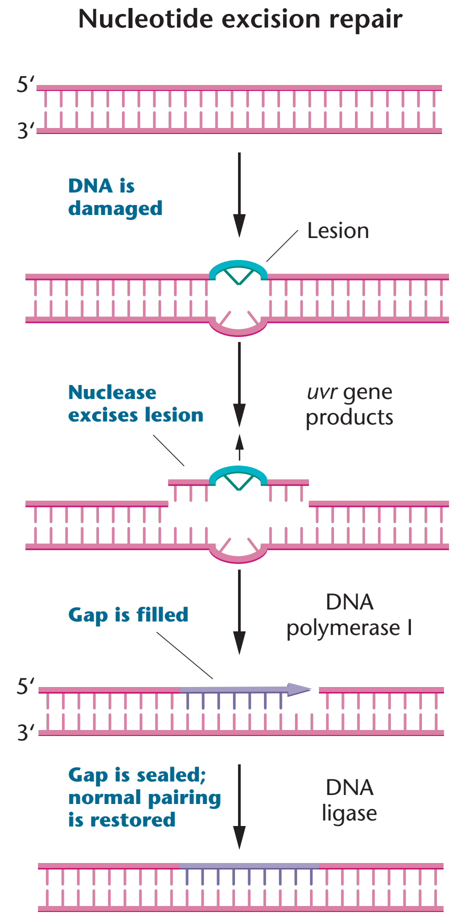


FIGURE 15.14 Nucleotide excision repair (NER) of a UV-induced thymine dimer. During repair, 13 nucleotides are excised in bacteria, and 28 nucleotides are excised in eukaryotes.

BER. The undamaged strand opposite the lesion is used as a template for the replication, resulting in repair.

Nucleotide Excision Repair and Xeroderma Pigmentosum in Humans

The mechanism of NER in eukaryotes is much more complicated than that in bacteria and involves many more proteins, encoded by about 30 genes. Much of what is known about the system in humans has come from detailed studies of individuals with **xeroderma pigmentosum (XP),** a rare recessive genetic disorder that predisposes individuals to severe skin abnormalities, skin cancers, and a wide range of other symptoms including developmental and neurological defects. Patients with XP are extremely sensitive to UV radiation in sunlight. In addition, they have a 2000-fold higher rate of cancer, particularly skin cancer, than the general population. The condition is severe and may be lethal, although early detection and protection from sunlight can arrest it (Figure 15.15).

The repair of UV-induced lesions in XP has been investigated *in vitro*, using human fibroblast cell cultures derived from normal individuals and those with XP. (Fibroblasts are undifferentiated connective tissue cells.) The results of these studies suggest that the XP phenotype is caused by defects in NER pathways and by mutations in more than one gene.

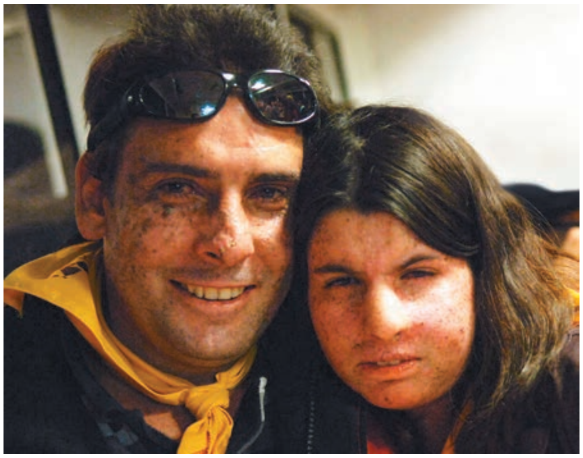


FIGURE 15.15 Two individuals with xeroderma pigmentosum. These XP patients show characteristic XP skin lesions induced by sunlight, as well as mottled redness (erythema) and irregular pigment changes to the skin, in response to cellular injury.

In 1968, James Cleaver showed that cells from XP patients were deficient in DNA synthesis other than that occurring during chromosome replication—a phenomenon known as **unscheduled DNA synthesis.** Unscheduled DNA synthesis is elicited in normal cells by UV radiation. Because this type of synthesis is thought to represent the activity of DNA polymerization during NER, the lack of unscheduled DNA synthesis in XP patients suggested that XP might be a deficiency in NER.

The involvement of multiple genes in NER and XP was further investigated by studies using **somatic cell hybridization.** Fibroblast cells from any two unrelated XP patients, when grown together in tissue culture, can fuse together, forming heterokaryons. A **heterokaryon** is a single cell with two nuclei from different organisms but a common cytoplasm. NER in the heterokaryon can be measured by the level of unscheduled DNA synthesis. If the mutation in each of the two XP cells occurs in the same gene, the heterokaryon, like the cells that fused to form it, will still be unable to undergo NER. This is because there is no normal copy of the relevant gene present in the heterokaryon.

However, if NER does occur in the heterokaryon, the mutations in the two XP cells must have been present in two different genes. Hence, the two mutants are said to demonstrate **complementation,** a concept discussed earlier (see Chapter 4). Complementation occurs because the heterokaryon has at least one normal copy of each gene in the fused cell. By fusing XP cells from a large number of XP patients, researchers were able to determine how many genes contribute to the XP phenotype. Based on these and other studies, XP patients were divided into seven complementation groups, indicating that at least seven different genes code for proteins that are involved in nucleotide excision repair in humans. A gene representing each of these complementation groups, *XPA* to *XPG* (*X*eroderma *P*igmentosum gene *A* to *G*), has now been identified, and a homologous gene for each has been identified in yeast.

Approximately 20 percent of XP patients do not fall into any of the seven complementation groups. Cells from most of these patients have mutations in the gene coding for DNA polymerase eta (Pol η), which is a lower-fidelity DNA polymerase that allows DNA replication to proceed past damaged DNA. Approximately another 6 percent of XP patients do not have mutations in either the seven complementation group genes or the *DNA polymerase eta* (*POLH*) gene, suggesting that other genes or mutations outside of coding regions may be involved in XP.

As a result of the study of defective genes in XP, a great deal is now known about how NER counteracts DNA damage in normal cells. The first step in humans is recognition of the damaged DNA by proteins encoded by the *XPC, XPE,* and *XPA* genes. These proteins then recruit the remainder of the repair proteins to the site of DNA damage. The *XPB* and *XPD* genes encode helicases, and the *XPF* and *XPG* genes encode nucleases. These and other factors form an excision repair complex that removes an approximately 28-nucleotide-long fragment from a DNA strand that contains the lesion. The resulting gap is then filled by DNA polymerase and sealed by DNA ligase.

Two other rare autosomal recessive diseases are associated with defects in NER pathways—Cockayne syndrome (CS) and trichothiodystrophy (TTD). The symptoms of CS include developmental and neurological defects and sensitivity to sunlight, but not an increase in cancers. Patients with CS age prematurely and usually die before the age of 20. The symptoms of TTD include dwarfism, intellectual disabilities, brittle skin and hair, and facial deformities. Like CS, these patients are sensitive to sunlight but do not have higher than normal rates of cancer. TTD patients have a median life span of six years.

Both CS and TTD arise from mutations in some of the same genes involved in XP (such as *XPB* and *XPD*), as well as other genes that encode proteins involved in NER within transcribed regions of the genome. It is not known why such a wide variety of different symptoms result from mutations in the same genes or DNA repair pathways; however, it may reflect the fact that products of many NER genes are also involved in other essential processes.

Double-Strand Break Repair in Eukaryotes

Thus far, we have discussed repair pathways that deal with damage or errors within one strand of DNA. We conclude our discussion of DNA repair by considering what happens in eukaryotic cells when both strands of the DNA helix are cleaved—as a result of exposure to ionizing radiation, for example. These types of damage are extremely dangerous to cells, leading to chromosome rearrangements, cancer, or cell death.

Specialized forms of DNA repair, the DNA **double-strand break (DSB) repair** pathways, are activated and are responsible for reattaching two broken DNA strands. Recently, interest in DSB repair has grown because defects in these pathways are associated with X-ray hypersensitivity and immune deficiency. Such defects may also underlie familial disposition to breast and ovarian cancer. Several human disease syndromes, such as Fanconi anemia and ataxia telangiectasia, result from defects in DSB repair.

One pathway involved in double-strand break repair is **homologous recombination repair.** The first step in this process involves the activity of an enzyme that recognizes the double-strand break and then digests back the 5′ ends of the broken DNA helix, leaving overhanging 3′ ends (**Figure 15.16**). One overhanging end searches for a region of sequence complementarity on the sister chromatid and then invades the homologous DNA duplex, aligning the complementary sequences.

Once aligned, DNA synthesis proceeds from the 3′ overhanging ends, using the undamaged homologous DNA strands as templates. The interaction of two sister chromatids is necessary because, when both strands of one helix are broken, there is no undamaged parental DNA strand available to use as a template DNA sequence during repair. After DNA repair synthesis, the resulting heteroduplex molecule is resolved and the two chromatids separate, as discussed earlier (see Chapter 11).

DSB repair usually occurs during the late S or early G2 phase of the cell cycle, after DNA replication, a time when sister chromatids are available to be used as repair templates. Because an undamaged template is used during repair synthesis, homologous recombination repair is an accurate process.

A second pathway, called **nonhomologous end joining,** also repairs double-strand breaks. However, as the name implies, the mechanism does not recruit a homologous region of DNA during repair. This system is activated in G1, prior to DNA replication. End joining involves a complex of many proteins and may include the DNA-dependent protein kinase and the breast cancer susceptibility gene product, BRCA1. These and other proteins bind to the free ends of the broken DNA, trim the ends, and ligate them back together. Because some nucleotide sequences are lost in the process of end joining, it is an error-prone repair system. In addition, if more than one chromosome suffers a double-strand break, the wrong ends could be joined together, leading to abnormal chromosome structures, such as those discussed earlier (see Chapter 8).

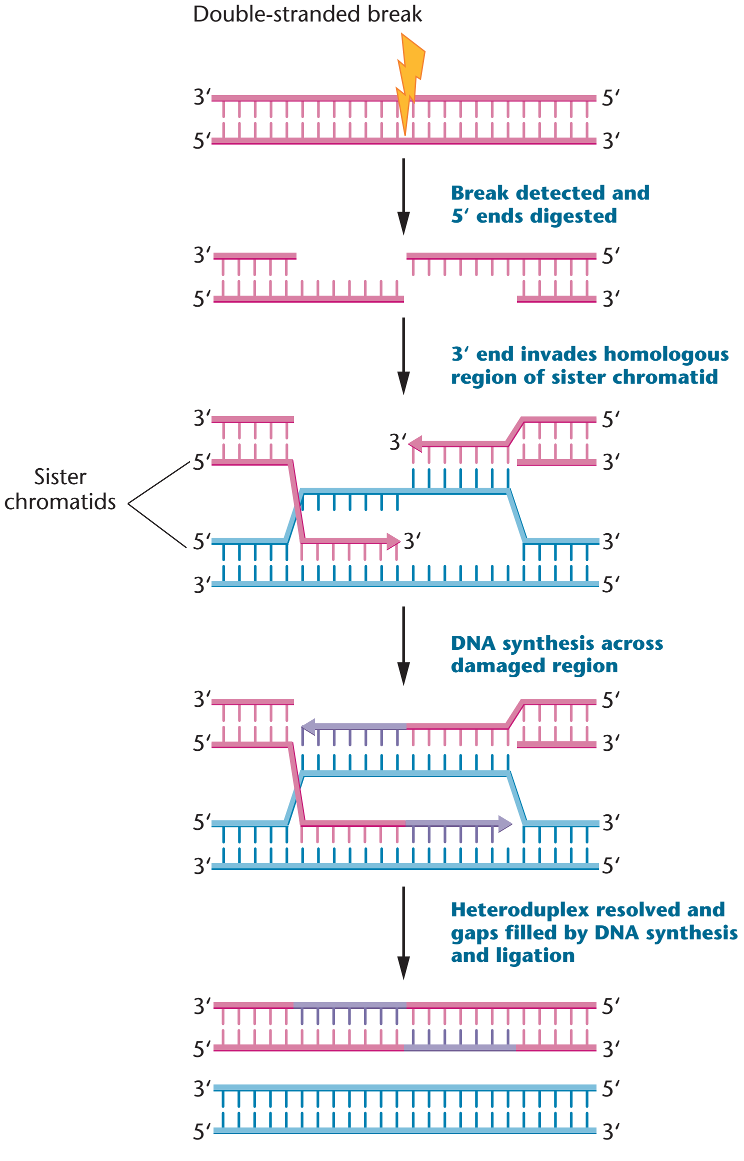

Double-stranded break
3′
5′
5′
3′
Break detected and 5′ ends digested
3′ end invades homologous region of sister chromatid
Sister chromatids
DNA synthesis across damaged region
Heteroduplex resolved and gaps filled by DNA synthesis and ligation

FIGURE 15.16 Steps in homologous recombination repair of double-stranded breaks.

NOW SOLVE THIS

15.4 Geneticists often use the alkylating agent ethylmethane sulfonate (EMS; see Figure 15.6) to induce mutations in *Drosophila*. Why is EMS a mutagen of choice for genetic research? What would be the effects of EMS in a strain of *Drosophila* lacking functional mismatch repair systems?

- **Hint:** *This problem asks you to evaluate EMS as a useful mutagen and to determine its effects in the absence of DNA repair. The key to its solution is to consider the chemical effects of EMS on DNA. Also, consider the types of DNA repair that may operate on EMS-mutated DNA and the efficiency of these processes.*

15.7 The Ames Test Is Used to Assess the Mutagenicity of Compounds

There is great concern about the possible mutagenic properties of any chemical that enters the human body, whether through the skin, the digestive system, or the respiratory tract. Examples of synthetic chemicals that concern us are those found in air and water pollution, food preservatives, artificial sweeteners, herbicides, pesticides, and pharmaceutical products. Mutagenicity can be tested in various organisms, including fungi, plants, and cultured mammalian cells; however, one of the most common tests, which we describe here, uses bacteria.

The **Ames test** (named for American biochemist Bruce Ames, who invented the assay in the 1960s) uses a number of different strains of the bacterium *Salmonella typhimurium* that have been selected for their ability to reveal the presence of specific types of mutations. For example, some strains are used to detect base-pair substitutions, and other strains detect various frameshift mutations. Each strain contains a mutation in one of the genes of the histidine operon. The mutant strains are unable to synthesize histidine (*his*$^-$ strains) and therefore require histidine for growth. The assay measures the frequency of reverse mutations that occur within the mutant gene, yielding wild-type bacteria (*his*$^+$ revertants) (Figure 15.17). The *his*$^-$ strains also have an increased sensitivity to mutagens due to the presence of mutations in genes involved in both DNA damage repair and the synthesis of the lipopolysaccharide barrier that coats these bacteria and protects them from external substances.

Many substances entering the human body are relatively innocuous until activated metabolically, usually in the liver, to more chemically reactive products. Thus, the Ames test includes a step in which the test compound is incubated *in vitro* in the presence of a mammalian liver extract. Alternatively, test compounds may be injected into a mouse where they are modified by liver enzymes and then recovered for use in the Ames test.

In the initial use of Ames testing in the 1970s, a large number of known **carcinogens,** or cancer-causing agents, were examined, and more than 80 percent of these were shown to be strong mutagens. This is not surprising, as the transformation of cells to the malignant state occurs as a result of mutations. For example, more than 60 compounds found in cigarette smoke test positive in the Ames test and cause cancer in animal tests. Although a positive response

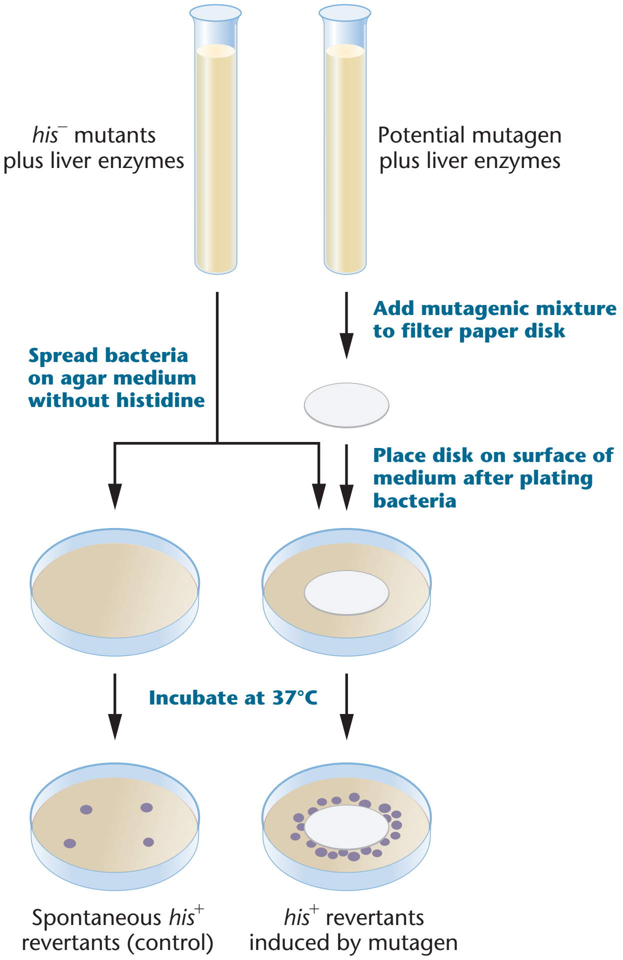

his⁻ mutants
plus liver enzymes
Potential mutagen
plus liver enzymes
Add mutagenic mixture to filter paper disk
Spread bacteria on agar medium without histidine
Place disk on surface of medium after plating bacteria
Incubate at 37°C
Spontaneous *his*⁺ revertants (control)
his⁺ revertants induced by mutagen

FIGURE 15.17 The Ames test, which screens compounds for potential mutagenicity. The high number of *his*$^+$ revertant colonies on the right side of the figure confirms that the substance being tested was indeed mutagenic.

in the Ames test does not prove that a compound is carcinogenic, the Ames test is useful as a preliminary screening device, as it is a rapid, convenient way to assess mutagenicity. Other tests of potential mutagens and carcinogens use laboratory animals such as rats and mice; however, these tests can take several years to complete and are more expensive. The Ames test is used extensively during the development of industrial and pharmaceutical chemical compounds.

15.8 Transposable Elements Move within the Genome and May Create Mutations

Transposable elements (TEs), informally known as "jumping genes," are DNA sequences that can move or transpose within and between chromosomes, inserting themselves into various locations within the genome. They can range from 50 to 10,000 base pairs in length.

TEs are present in the genomes of all organisms from bacteria to humans. Not only are they ubiquitous, but they also make up large portions of some eukaryotic genomes. For example, almost 50 percent of the human genome is derived from TEs. Some organisms with unusually large genomes, such as salamanders and barley, contain hundreds of thousands of copies of various types of TEs constituting as much as 85 percent of these genomes.

Although the possible functions of these elements are uncertain, data from human genome sequencing suggest that some genes may have evolved from TEs and that the presence of these elements may help to modify and reshape the genome. In addition, the movement of TEs from one place in the genome to another has the capacity to disrupt genes and cause mutations, as well as to create chromosomal damage such as double-strand breaks. TEs also act as sites of genome rearrangement events, when homologous recombination occurs between DNA sequences with sequence similarities.

Since their discovery, TEs have also become valuable tools in genetic research. Geneticists harness these DNA elements as mutagens, as cloning tags, and as vehicles for introducing foreign DNA into model organisms. One use of TEs in genetic research is described in the Modern Approaches to Understanding Gene Function feature on page 366.

TEs can be classified into two groups, based on their methods of transposition. *Retrotransposons* move using an RNA intermediate, and *DNA transposons* move in and out of the genome as DNA elements. We will look at both groups in the sections that follow.

DNA Transposons

DNA transposons move from one location to another without going through an RNA intermediate stage. They are abundant in many organisms from bacteria to humans (**Table 15.5**).

DNA transposons share several structural features that are important for their function (**Figure 15.18**). Inverted terminal repeats (ITRs) are located on each end of the TE, and an open reading frame (ORF) codes for the enzyme transposase; both are required for movement of the TE in and out of the

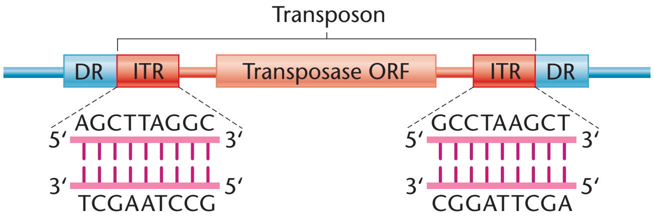


FIGURE 15.18 Structural features of DNA transposons. DNA transposons, shown in red, contain an open reading frame (ORF) that encodes the enzyme transposase. Some DNA transposons also contain ORFs encoding other proteins in addition to transposase. Inverted terminal repeats (ITRs), shown in detail below the main diagram, are short DNA sequences that are inverted relative to each other. Direct repeats (DRs, shown in blue) flank the DNA transposon in the chromosomal DNA.

TABLE 15.5 Examples of DNA Transposons

Category	Element	Example	Organism
Autonomous	Insertion sequences	IS*1*	Bacteria
	Composite elements	Tn*5*, Tn*10*	Bacteria
	Activator element	*Ac*	Maize
	Full-length *P* elements	*P* element	*Drosophila*
Nonautonomous	*Dissociation* element	*Ds*	Maize
	Miniature inverted repeat TEs (MITEs)	*MiHsmar1*	Plants, vertebrates
	Mariner-like	*Tc1*	Fish, other vertebrates
	P elements with internal deletions	*P* element	*Drosophila*

genome. ITRs are DNA sequences of between 9 and 40 bp long that are identical in sequence, but inverted relative to each other. ITRs are essential for transposition and are recognized and bound by the transposase enzyme. Short direct repeats (DRs) are present in the host DNA, flanking each TE insertion. These flanking DRs are created as a consequence of the TE insertion process, as described later in this section.

DNA transposons vary considerably in length and are classified as either autonomous or nonautonomous (see Table 15.5). *Autonomous transposons* are able to transpose by themselves, as they encode a functional transposase enzyme and have intact ITRs. *Nonautonomous transposons* cannot move on their own because they do not encode their own functional transposase enzyme. They require the presence of an autonomous transposon elsewhere in the genome, so that the transposase synthesized by the autonomous element can be used by the nonautonomous element for transposition.

Most DNA transposons move through the genome using "cut-and-paste" mechanisms, in which the transposon is physically cut out of the genome and then inserted into a new position in the same or a different chromosome. In the first step, the transposase enzyme binds to the ends of the TE, making cuts between the transposon DNA and chromosomal DNA (**Figure 15.19**). This step releases the

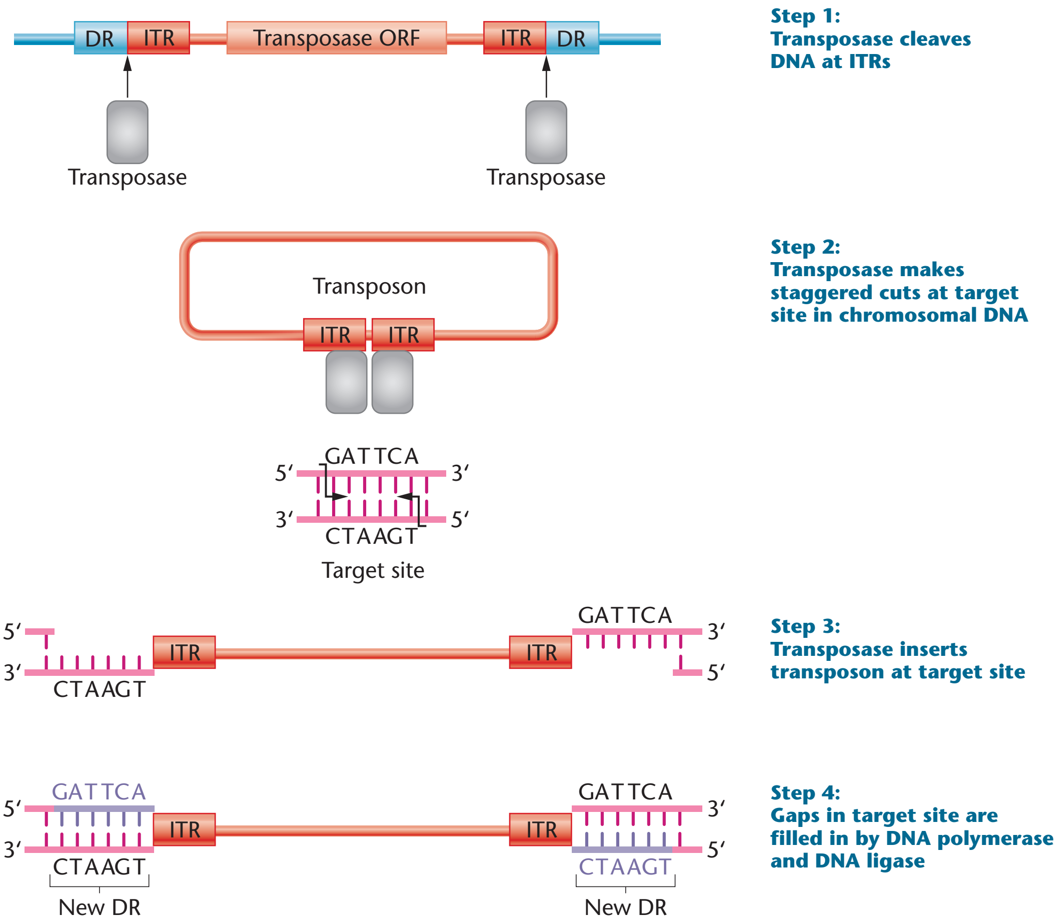

DR
ITR
Transposase ORF
ITR
DR
Transposase
Transposase
Step 1: Transposase cleaves DNA at ITRs
Transposon
ITR
ITR
Step 2: Transposase makes staggered cuts at target site in chromosomal DNA
GATTCA
5′
3′
3′
5′
CTAAGT
Target site
5′
3′
CTAAGT
ITR
ITR
GATTCA
3′
5′
Step 3: Transposase inserts transposon at target site
GATTCA
5′
3′
ITR
ITR
GATTCA
3′
5′
CTAAGT
CTAAGT
New DR
New DR
Step 4: Gaps in target site are filled in by DNA polymerase and DNA ligase

FIGURE 15.19 Steps in DNA transposon transposition.

transposon, which remains bound by transposase. In the second step, the transposase enzyme makes staggered cuts at a new location in the chromosomal DNA, leaving 5′ and 3′ single-stranded DNA overhangs. Transposase then inserts the transposon into this new location (Step 3). In the final step, the 5′ and 3′ DNA overhangs are filled in by DNA polymerase and sealed by DNA ligase. In this way, a new DR is created, flanking the new insertion. Usually, but not always, the site from which the DNA transposon was cut is repaired accurately, leaving no trace of the original DNA transposon.

Examples of DNA transposons and the ways in which their movements can affect gene expression are described next.

DNA Transposons—the *Ac–Ds* System in Maize

DNA transposons were first discovered by Barbara McClintock in the late 1940s as a result of her research on the genetics of maize. Her work involved analysis of the genetic behavior of two mutations, ***Dissociation (Ds)*** and ***Activator (Ac)***, expressed in either the endosperm or aleurone layers of maize seeds. She then correlated her genetic observations with cytological examinations of maize chromosomes. Initially, McClintock determined that *Ds* was located on chromosome 9. If *Ac* was also present in the genome, *Ds* induced breakage at a point on the chromosome adjacent to its own location. If chromosome breakage occurred in somatic cells during their development, progeny cells often lost part of the broken chromosome, causing a variety of phenotypic effects. The chapter-opening photo illustrates the types of phenotypic effects caused by *Ds* mutations in kernels of corn.

Subsequent analysis suggested to McClintock that both *Ds* and *Ac* elements sometimes moved to new chromosomal locations. While *Ds* moved only if *Ac* was also present, *Ac* was capable of autonomous movement. Where *Ds* came to reside determined its genetic effects—that is, it might cause chromosome breakage, or it might inhibit expression of a certain gene. In cells in which *Ds* caused a gene mutation, *Ds* might move again, restoring the gene mutation to wild type. Figure 15.20 illustrates the types of movements and effects brought about by *Ds* and *Ac* elements.

In McClintock's original observation, pigment synthesis was restored in cells in which the *Ds* element jumped out of chromosome 9. McClintock concluded that *Ds* was a **mobile controlling element.** Similar mobility was later also revealed for *Ac*. We now commonly refer to these as transposable elements (TEs).

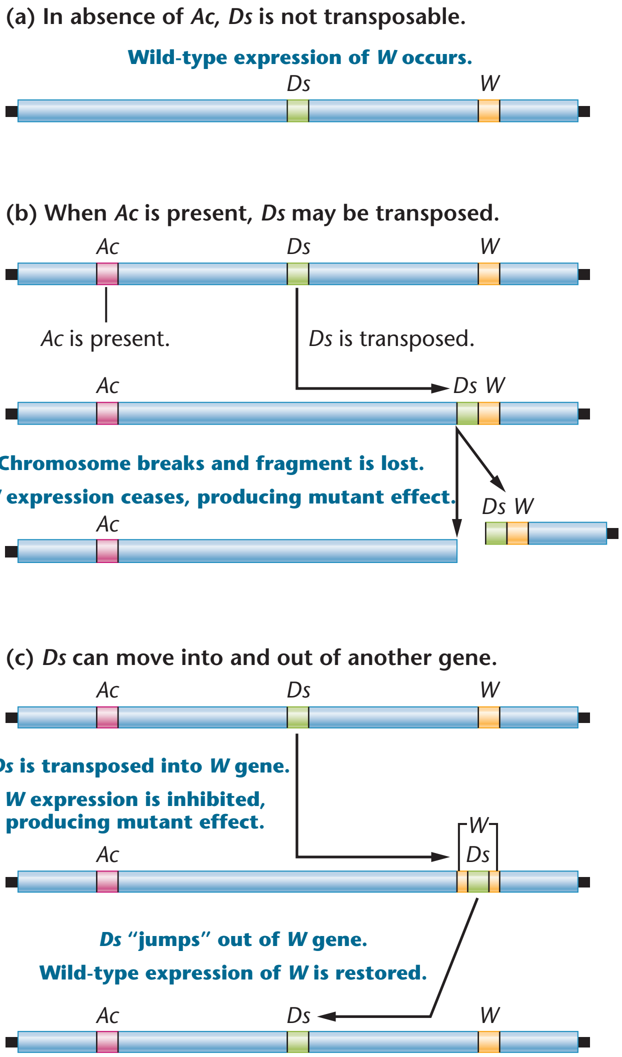

(a) In absence of Ac, Ds is not transposable.
Wild-type expression of W occurs.
Ds
W
(b) When Ac is present, Ds may be transposed.
Ac
Ds
W
Ac is present.
Ds is transposed.
Ac
Ds W
Chromosome breaks and fragment is lost.
W expression ceases, producing mutant effect.
Ac
Ds W
(c) Ds can move into and out of another gene.
Ac
Ds
W
Ds is transposed into W gene.
W expression is inhibited, producing mutant effect.
Ac
W
Ds
Ds "jumps" out of W gene.
Wild-type expression of W is restored.
Ac
Ds
W

FIGURE 15.20 Effects of *Ac* and *Ds* elements on gene expression. (a) If *Ds* is present in the absence of *Ac*, there is normal expression of a distantly located hypothetical gene *W*. (b) In the presence of *Ac*, *Ds* may transpose to a region adjacent to *W*. *Ds* can induce chromosome breakage, which may lead to loss of a chromosome fragment bearing the *W* gene. (c) In the presence of *Ac*, *Ds* may transpose into the *W* gene, disrupting *W*-gene expression. If *Ds* subsequently transposes out of the *W* gene, *W*-gene expression may return to normal.

Several *Ac* and *Ds* elements have now been analyzed, and the relationship between the two elements has been clarified. The first *Ds* element studied (*Ds*9) is nearly identical to *Ac* except for a 194-bp deletion within the transposase gene. The deletion of part of the transposase gene in the *Ds*9 element explains its dependence on the *Ac* element for transposition. Several other *Ds* elements have also been sequenced, and each contains an even larger deletion within the transposase gene. In each case, however, the ITRs are retained.

Although the significance of Barbara McClintock's mobile controlling elements was not fully appreciated following her initial observations, molecular analysis has since verified her conclusions. She was awarded the Nobel Prize in Physiology or Medicine in 1983.

Retrotransposons

Retrotransposons are TEs that amplify and move within the genome using RNA as an intermediate. Their methods of transposition are sometimes described as "copy-and-paste" mechanisms. In many ways, retrotransposons resemble retroviruses, which replicate using similar mechanisms. However, retrotransposons do not encode all of the proteins that are required to form mature virus particles and therefore are not infective.

Retrotransposons can be very abundant in some organisms. For example, maize genomes are made up of as much as 78 percent retrotransposon DNA. Approximately 42 percent of the human genome consists of retrotransposons or their remnants, whereas only approximately 3 percent of the human genome consists of DNA transposons.

There are two types of retrotransposons—the long-terminal-repeat (LTR) retrotransposons and the non-LTR retrotransposons. In addition, retrotransposons, like DNA transposons, can be either autonomous or nonautonomous. Examples of the various types of TEs are listed in **Table 15.6**.

Like DNA transposons, retrotransposons encode proteins that are required for their transposition and are flanked by direct repeats at their insertion sites. The structures of the two types of retrotransposons are shown in **Figure 15.21**.

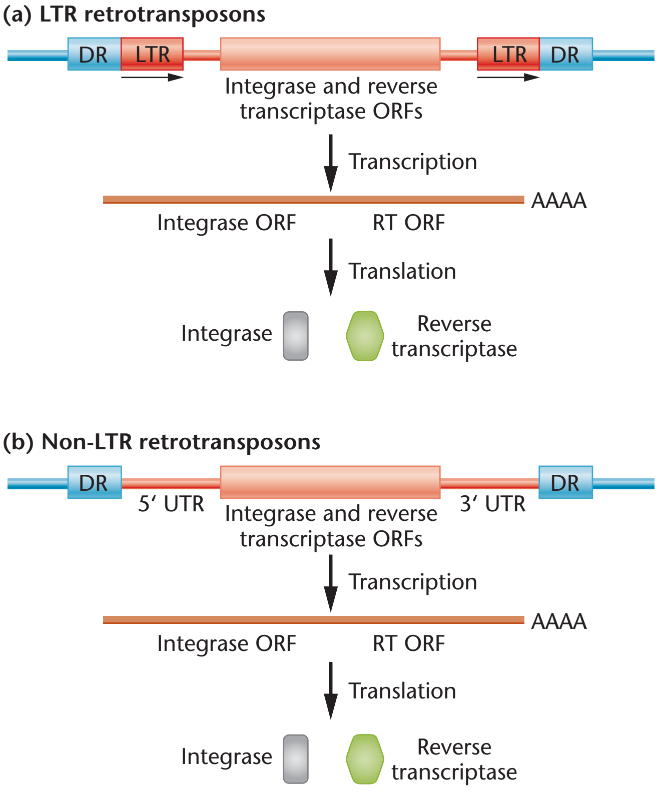


FIGURE 15.21 Retrotransposon structures. (a) LTR retrotransposons, shown in red, contain open reading frames (ORFs) that encode the enzymes integrase and reverse transcriptase (RT). Transcription promoters and polyadenylation sites are located within 5′ and 3′ long terminal repeats (LTRs). The bottom part of the diagram shows transcription of the LTR retrotransposon and translation into integrase and reverse transcriptase. (b) Non-LTR retrotransposons also encode integrase and reverse transcriptase, but are lacking LTRs. Transcription promoters and polyadenylation sites are located within 5′ and 3′ untranslated regions (UTRs).

The steps in retrotransposon transposition involve the actions of both retrotransposon-encoded proteins and those that are part of the cell's normal transcriptional and translational machinery (**Figure 15.22**). In the first step, the cell's RNA polymerases transcribe the retrotransposon DNA into one or more RNA copies. In the second step, the RNA copies are translated into the two enzymes required for transposition—reverse transcriptase and integrase. The retrotransposon RNAs are then converted to double-stranded DNA copies through the actions of reverse transcriptase. The ends of the double-stranded DNAs are recognized by integrase, which then inserts the retrotransposons back into the genome. Because many RNA copies can be converted to DNA and transposed in this way, retrotransposons are able to

TABLE 15.6 Examples of retrotransposons

Category	Element	Example	Organism
LTR, autonomous	Ty1-*copia* group	*copia*	Plants, animals, algae
	Ty1-*gypsy* group	*gypsy*	Plants, animals, fungi
LTR, nonautonomous	Large retrotransposon derivatives (LARDS)	*Dasheng*	plants
Non-LTR, autonomous	LINE elements	*L1*	Humans, other organisms
	I elements	*I factor*	*Drosophila*
Non-LTR, nonautonomous	SINE elements	*Alu*	Humans, other organisms
	SINE-VNTR-Alu elements	*SVA*	Humans

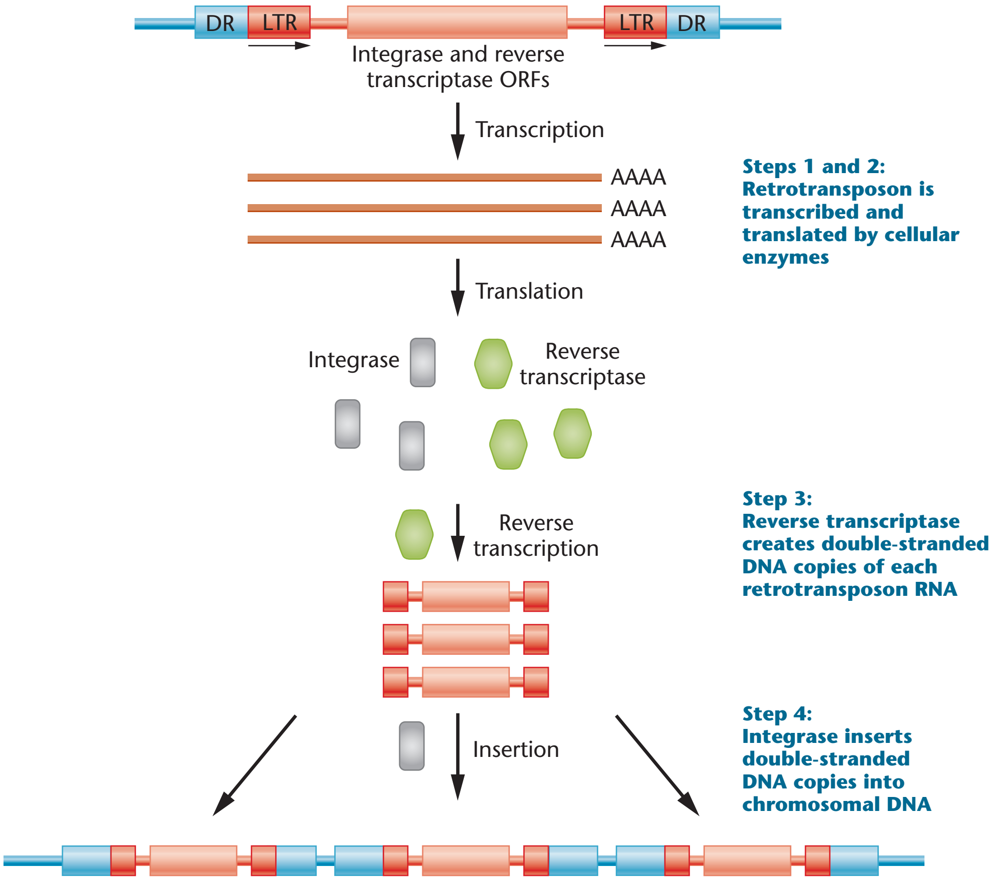

DR
LTR
LTR
DR
Integrase and reverse transcriptase ORFs
Transcription
AAAA
AAAA
AAAA
Steps 1 and 2: Retrotransposon is transcribed and translated by cellular enzymes
Translation
Integrase
Reverse transcriptase
Reverse transcription
Step 3: Reverse transcriptase creates double-stranded DNA copies of each retrotransposon RNA
Insertion
Step 4: Integrase inserts double-stranded DNA copies into chromosomal DNA

FIGURE 15.22 Steps in retrotransposon transposition.

accumulate rapidly and may create mutations at many sites in the genome. In addition, the original retrotransposon is not excised during transposition.

Next, we will look at one well-studied example of a retrotransposon—*copia*—and describe its effects on the *white* locus in *Drosophila*.

Retrotransposons—the *Copia*–*White-Apricot* System in *Drosophila*

In 1975, David Hogness and his colleagues David Finnegan, Gerald Rubin, and Michael Young identified a class of retrotransposons in *Drosophila melanogaster* that they designated as ***copia***. These elements are transcribed into "copious" amounts of RNA (hence their name). *Copia* elements are present in 10 to 100 copies in the genomes of *Drosophila* cells. Mapping studies show that they are transposable to different chromosomal locations and are dispersed throughout the genome.

Each *copia* element consists of approximately 5000 to 8000 bp of DNA, including a **long terminal repeat (LTR)** sequence of 267 bp at each end. Like other LTR retrotransposons, transcription of the *copia* element begins in the 5′ LTR, which contains a promoter and transcription start site. The transcript is cleaved and polyadenylated within the 3′ LTR, which contains a polyadenylation site. These features allow the retrotransposon to be transcribed by the cell's RNA polymerase.

Insertion of *copia* is dependent on the presence of the LTR sequences and seems to occur preferentially at specific target sites in the genome. *Copia* elements confer regulatory effects at the point of their insertion in the chromosome. Certain mutations, including those affecting eye color and segment formation, are due to *copia* insertions within genes.

One of the earliest descriptions of *copia* effects came from research into the *white-apricot* mutation in *Drosophila*. This mutation first appeared spontaneously in 1923 and changed the *Drosophila* eye color from a wild-type red to an orange-yellow color [Figure 15.23(a)]. Later DNA sequencing studies demonstrated that the mutation was caused by an insertion of *copia* into the second intron of the *white* gene. As a result of this insertion, most of the transcripts that originate from the *white* gene promoter terminate prematurely within the 3′ LTR of the *copia* retrotransposon. These prematurely terminated transcripts do not encode functional *white* gene product, resulting in a loss of red pigment in the eye [Figure 15.23(b)]. Because some *white* gene transcripts read through the *copia* element, enough white gene product is produced to yield a light-orange colored eye.

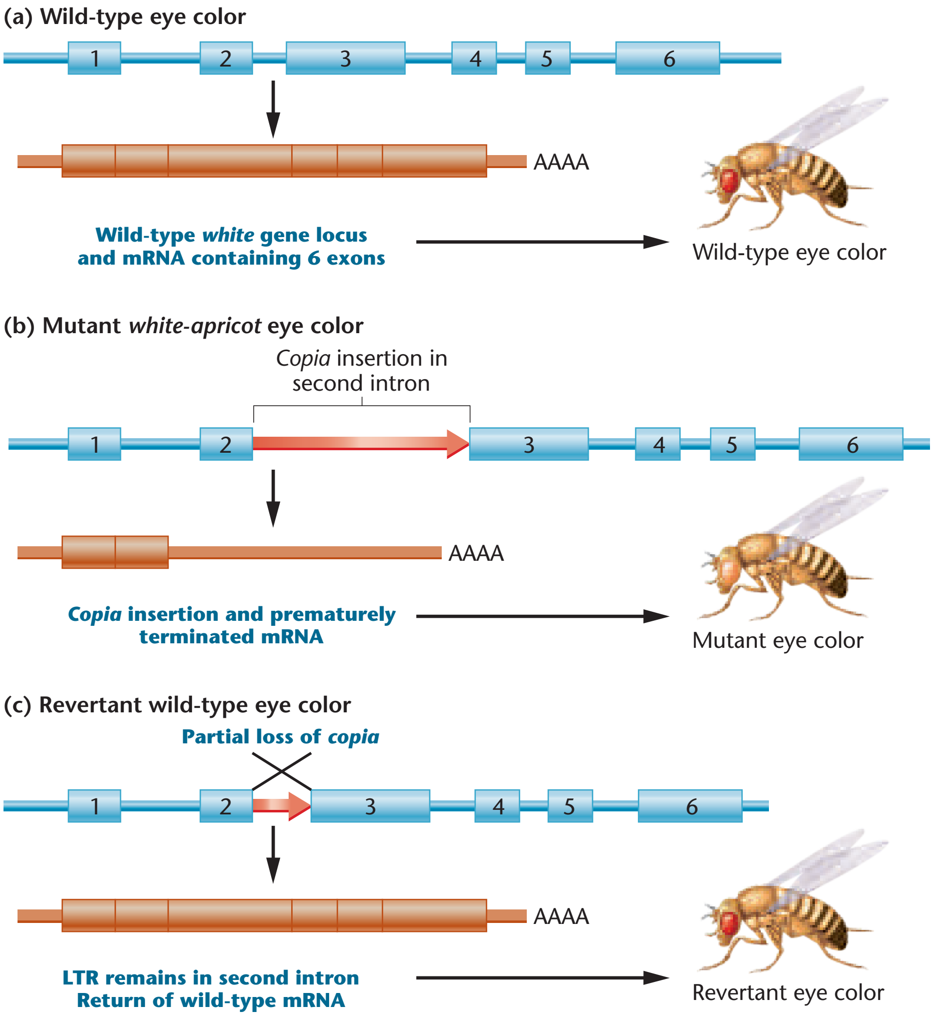

(a) Wild-type eye color
1
2
3
4
5
6
AAAA
Wild-type *white* gene locus and mRNA containing 6 exons
Wild-type eye color
(b) Mutant *white-apricot* eye color
Copia insertion in second intron
1
2
3
4
5
6
AAAA
Copia insertion and prematurely terminated mRNA
Mutant eye color
(c) Revertant wild-type eye color
Partial loss of *copia*
1
2
3
4
5
6
AAAA
LTR remains in second intron
Return of wild-type mRNA
Revertant eye color

FIGURE 15.23 Effects of *copia* insertion into the *white* gene of *Drosophila*. (a) The *white* gene (top, blue) in wild-type *Drosophila* contains six exons, all of which are present in the mRNA (bottom, orange). (b) The *white* gene in mutant *white-apricot Drosophila* contains an insertion of *copia* (red) in the second intron and a prematurely terminated mRNA containing only two exons. (c) Revertant *Drosophila* have lost most of the *copia* element from the second intron, resulting in wild-type mRNA and wild-type eye color.

A partial revertant of the *white-apricot* mutation has been characterized [Figure 15.23(c)]. This revertant has an eye color that is almost wild type. Sequence analysis shows that most of the *copia* element in the second intron is absent, with only one LTR remaining. This loss of *copia* is believed to be due to homologous recombination between the two LTR sequences, resulting in a DNA deletion.

Transposable Elements in Humans

Recent genomic sequencing data reveal that almost half of the human genome is composed of TE DNA. As discussed earlier (see Chapter 12), the major families of human TEs are the **long interspersed elements (LINEs)** and **short interspersed elements (SINEs)**, both of which are non-LTR retrotransposons. Together, they make up 34 percent of the human genome. Other families of TEs account for a further 11 percent (Table 15.7). As coding sequences comprise only about 1 percent of the human genome, there is about 40 to 50 times more TE DNA in the human genome than DNA in functional genes.

Although most human TEs appear to be inactive, the potential mobility and mutagenic effects of these elements have far-reaching implications for human genetics, as can be seen in an example of a TE "caught in the act." The case involves a male child with hemophilia. One cause of hemophilia is a defect in blood-clotting factor VIII, the product of an X-linked gene. Haig Kazazian and colleagues found LINEs inserted at two points within the gene. Researchers were interested in determining if one of the mother's X chromosomes also contained this specific LINE. If so, the unaffected mother would be heterozygous and could pass the LINE-containing chromosome to her son. The surprising finding was that the LINE sequence

TABLE 15.7 Transposable Elements in the Human Genome

Element Type	Length	Copies in Genome	% of Genome
LINEs	1–6 kb	850,000	21
SINEs	100–500 bp	1,500,000	13
LTR elements	<5 kb	443,000	8
DNA transposons	80–300 bp	294,000	3
Unclassified	—	3,000	0.1

was *not* present on either of her X chromosomes but *was* detected on chromosome 22 of both parents. This suggests that this mobile element may have transposed from one chromosome to another in the gamete-forming cells of the mother, prior to being transmitted to the son.

LINE insertions into the human *dystrophin* gene (another X-linked gene) have resulted in at least two separate cases of Duchenne muscular dystrophy. In one case, a LINE inserted into exon 48, and in another case, it inserted into exon 44, both leading to frameshift mutations and premature termination of translation of the dystrophin protein. There are also reports that LINEs have inserted into the *APC* and *c-myc* genes, leading to mutations that may have contributed to the development of some colon and breast cancers. In the latter cases, the transposition had occurred within one or a few somatic cells. As of 2009, researchers have determined that cases of at least 11 human diseases are due to insertions of LINE elements.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Transposon-Mediated Mutations Reveal Genes Involved in Colorectal Cancer

In this chapter you learned about transposons and how they can move within a genome. Scientists have taken advantage of transposons, using them as a tool for mutagenesis of specific and random DNA sequences in the genome. This represents a **forward genetics** approach to create transgenic and knockout animals to learn about gene function.

Human colorectal cancers (CRCs) arise from complex underlying genetic and epigenetic alterations. Some genes are thought to be *drivers of tumorigenesis*; that is, they cause tumor formation. Other mutations are thought to be *passenger mutations*: They appear at some point during the progression of CRCs, but they have little to no direct effect on the origin of a tumor. To develop novel, targeted therapies for treating CRCs, it is essential that scientists distinguish driver mutations from passenger mutations. Here we profile a landmark study in which transposons were used to reveal CRC driver genes in a mouse model.

Researchers used a **conditional knockout** strategy involving a mating between different transgenic animals. One transgenic mouse strain contained a transposase gene coupled to a promoter sequence for the *Villin* gene, a gene that is expressed in the small intestine. As a result, these animals mainly express high levels of transposase in the epithelial cells of the GI tract. Another transgenic strain, called *T2/Onc*, contains a transposon. This transposon will not move in this parent strain, as it lacks the correct transposase. Crossing these mice resulted in offspring with active transposase confined to cells of the GI tract. The transposase allows the transposition of the *T2/Onc* transposon into other positions within the mouse genome. In theory, transposition mutagenesis of driver genes involved in CRCs would produce mice with mutant phenotypes resembling CRCs.

Results:

It was found that 72 percent of transgenic mice offspring expressing transposase in the GI tract had intestinal tumors (see photos). These animals died at a faster rate than control animals did and at a younger age. Analysis of intestinal lesions confirmed characteristics of CRC tumor types. DNA from tumors was sequenced to identify the locations of transposon insertions, and an analysis of over 16,000 transposons revealed 77 candidate genes for CRC. The genes revealed in this study were compared to a recent large-scale exon sequencing project that analyzed mutations in genes from human CRC tumor samples. Sixty of the mouse genes identified in this study were shown to be mutated or improperly expressed in human CRC. This transposon screen also revealed 17 candidate genes that had not been previously implicated in CRC.

Conclusions:

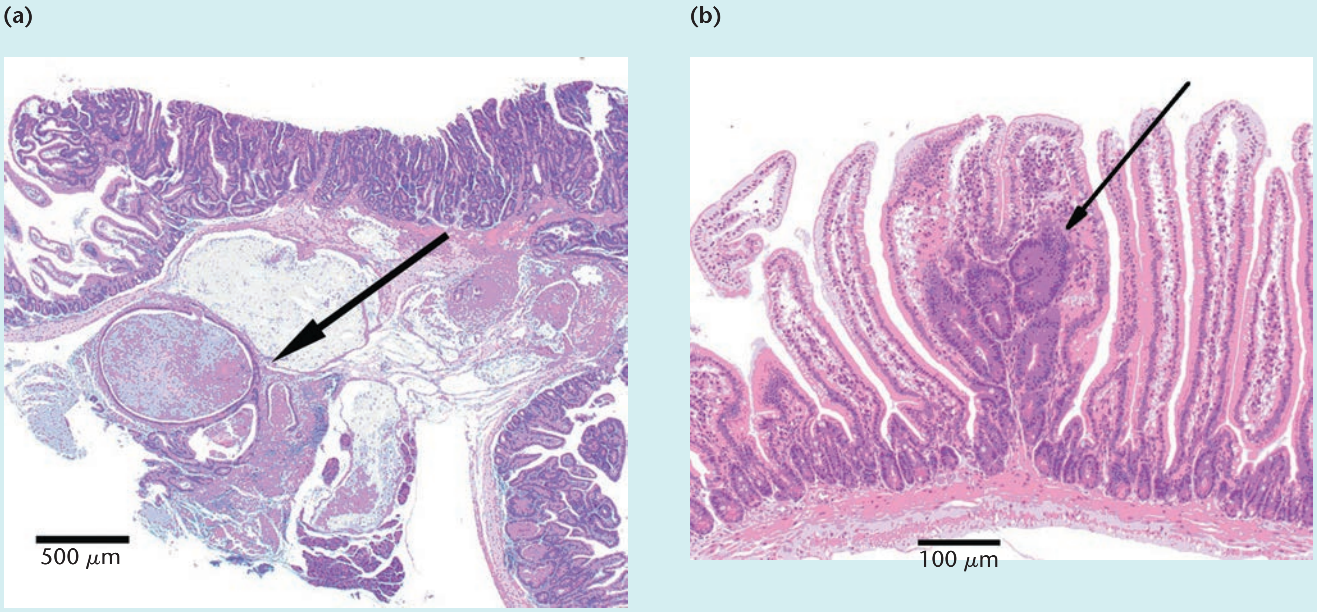
This transposon screen produced mice with identified mutations resulting in tumors of the GI tract and revealed candidate genes that may be drivers of CRC tumor formation in humans. It suggests that tumorigenesis of CRC is driven by a small subset of mutated genes. Discovering significant overlap between mouse candidate genes and human genes known to be involved in CRC suggests that this mouse model will be useful for developing drug-based and other strategies for treating CRC in humans.

Reference:

Starr, T. K., et al. (2009). A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* 323:1747–1750.

Questions to Consider:

1. The *T2/Onc* transposon system in this paper is sometimes called "Sleeping Beauty." Why do you think that is a good name? Do an Internet search to learn more about this system to help you answer this question.
2. What do you think are limitations to this type of screening approach?
3. *Apc*, *Tcf12,* and *Fbxw7* were among the top-ranked driver genes identified in this study based on the number of transposon insertions in transgenic mice and their mutant or aberrant expression in human CRCs. Search the Online Mendelian Inheritance in Man (OMIM) Web site (http://www.omim.org) to learn what is known about the function of these genes.



Photomicrographs of hematoxylin and eosin-stained tissue from the small intestine of transgenic animals showing examples of GI tumors. (a) an adenocarcinoma (arrow) *and (b) a cluster of tumors that has invaded the GI epithelium* (arrow).

SINE insertions are also responsible for more than 30 cases of human disease. In one case, an ***Alu* element** integrated into the *BRCA2* gene, inactivating this tumor-suppressor gene and leading to a familial case of breast cancer. Other genes that have been mutated by *Alu* integrations are the *factor IX* gene (leading to hemophilia B), the *ChE* gene (leading to acholinesterasemia), and the *NF1* gene (leading to neurofibromatosis).

Transposable Elements, Mutations, and Evolution

TEs can have a wide range of effects on genes, based on where they are inserted and their composition. Here are a few examples:

- The insertion of a TE into one of a gene's coding regions may disrupt the gene's normal translation reading frame or may induce premature termination of translation of the gene's mRNA.
- The insertion of a TE containing polyadenylation or transcription termination signals into a gene's intron may bring about termination of the gene's transcription within the element. In addition, it can cause aberrant splicing of an RNA transcribed from the gene.
- Insertions of a TE into a gene's transcription regulatory region may disrupt the gene's normal regulation or may cause the gene to be expressed differently as a result of the presence of the TE's own promoter or enhancer sequences.
- The presence of two or more identical TEs in a genome creates the potential for recombination between the transposons, leading to duplications, deletions, inversions, or chromosome translocations. Any of these rearrangements may bring about phenotypic changes or disease.

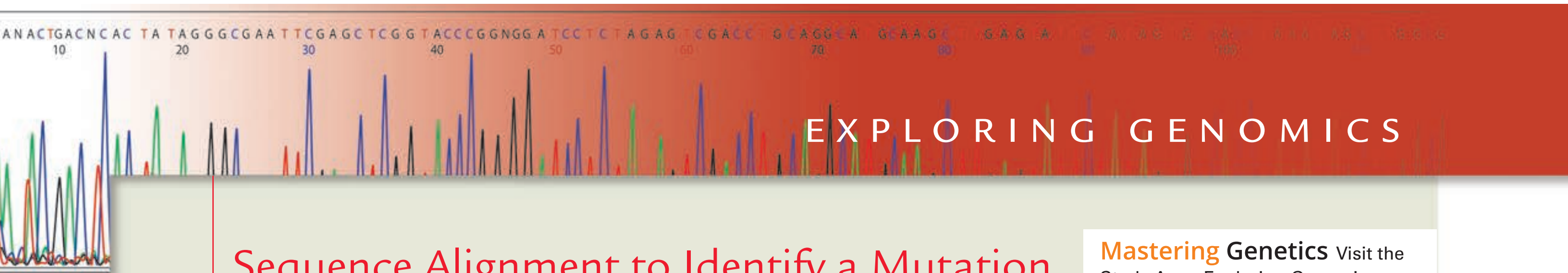
It is thought that about 0.2 percent of detectable human mutations may be due to TE insertions. Other organisms appear to suffer more damage due to transposition. For example, about 10 percent of new mouse mutations and 50 percent of *Drosophila* mutations are caused by insertions of TEs in or near genes.

Because of their ability to alter genes and chromosomes, TEs contribute to evolution. Some mutations caused by TE insertions or deletions may be beneficial to the organism under certain circumstances. These mutations may be selected for and maintained through evolution.

In some cases, TEs themselves may be modified to perform functions that become beneficial to the organism, and these may be selected for and maintained. One example of TEs that contributed to evolution is provided by *Drosophila* telomeres. LINE-like elements are present at the ends of *Drosophila* chromosomes, and these elements have evolved to act as telomeres, maintaining the length of *Drosophila* chromosomes over successive cell divisions.

Another example of evolved TEs are the *RAG1* and *RAG2* genes in humans. These genes encode **recombinase** enzymes that are essential to the development of the immune system, and are involved in rearrangements and recombinations of genes encoding immunoglobulin and T-cell receptors. These two genes appear to have evolved from TEs that entered ancestral vertebrate genomes more than 500 million years ago.

TEs may also affect the evolution of genomes by altering gene-expression patterns in ways that are subsequently retained by the host. For example, the human *amylase* gene contains an enhancer that causes the gene to be expressed in the parotid gland. This enhancer evolved from TE sequences that were inserted into the gene-regulatory region early in primate evolution.



EXPLORING GENOMICS

Sequence Alignment to Identify a Mutation

Mastering Genetics Visit the Study Area: Exploring Genomics

In this chapter, we examined the causes of different types of mutations and how mutations affect phenotype by altering the structure and function of proteins. The emergence of genomics, bioinformatics, and proteomics as key disciplines in modern genetics has provided geneticists with an unprecedented set of tools for identifying and analyzing mutations in gene and protein sequences.

In this exercise we will use the **ExPASy Bioinformatics Resource Portal**, a Web site hosted by the Swiss Institute for Bioinformatics that provides a wealth of resources for studying proteins. Here we will use an ExPASy program called SIM (for "similarity" in sequence) to compare two polypeptide sequences so as to pinpoint a mutation. Once the mutation has been identified, you will learn more about the gene encoding these polypeptides and about a human disease condition associated with this gene.

■ Exercise I — Identifying a Missense Mutation Affecting a Protein in Humans

1. Begin this exercise by accessing the ExPASy Web site at http://web.expasy.org/sim/. The SIM feature is an algorithm-based software program that allows us to compare multiple polypeptide sequences by looking for amino acid similarity in the sequences.
2. Go to the Study Area for *Concepts of Genetics*, and open the Exploring Genomics exercise for this chapter. At the Web site we provide amino acid sequences for polypeptides expressed in two different people (Person A and Person B). Note that the amino acid sequence is provided using the single-letter code for each amino acid (see the accompanying table for amino acid names and single-letter codes).
3. Copy and paste each sequence into separate "SEQUENCE" text boxes in SIM. Use the Person A sequence for sequence 1 and the Person B sequence for sequence 2. Click the "User-entered sequence" button for each. Name the sequences Person A and Person B as appropriate. Submit the sequences for comparison, and then answer the following questions:
   a. How many amino acids (called "residues" in the SIM results report) are in each polypeptide sequence that was analyzed?
   b. Look carefully at the alignment results. Can you find any differences in amino acid sequence when comparing these two polypeptides? What did you find?

■ Exercise II – Identifying the Genetic Basis for a Human Genetic Disease Condition

1. From the ExPASy Web site use the BLAST link (http://web.expasy.org/blast/) and run a protein BLAST (blastp) search to identify which polypeptide you have been studying. Explore the BLAST reports for the top three protein sequences that aligned with your query sequence by clicking on the link for each sequence. Pay particular attention to the "Comment" section of each report to help you answer the questions in Exercise 3.
2. Now that you know what gene you are working with, go to PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and search for a review article from the authors Vajo, Z., Francomano, C. A., and Wilkin, D. J.
3. Answer the following questions:
   a. What gene codes for the polypeptides have you been studying?
   b. What is the function of this protein?
   c. Based on what you learned from the alignment results you analyzed in Exercise I, the BLAST reports, and your PubMed search, what human disease is caused by the mutation you identified in Exercise I? Explain your answer, and briefly describe phenotypes associated with the disease.

CASE STUDY An Unexpected Diagnosis

Six months pregnant, an expectant mother had a routine ultrasound that showed that the limbs of the fetus were unusually short. Her physician suspected that the baby might have a genetic form of dwarfism called achondroplasia, an autosomal dominant trait occurring with a frequency of about 1 in 27,000 births. The parents were directed to a genetic counselor to discuss this diagnosis. In the conference, they learned that achondroplasia is caused by a mutant allele. Sometimes it is passed from one generation to another, but in 80 percent of all cases it is the result of a spontaneous mutation that arises in a gamete of one of the parents. They also learned that most children with achondroplasia have normal intelligence and a normal life span.

1. What information would be most relevant to concluding which of the two mutation origins, inherited or new, most likely pertains in this case? How does this conclusion impact on this couple's decision to have more children?
2. It has been suggested that prenatal genetic testing for achondroplasia be made available and offered to all women. Would you agree with this initiative? What ethical considerations would you consider when evaluating the medical and societal consequences of offering such testing?

For related reading see Radoi, V., et al. (2016). How to provide a genetic counseling in a simple case of antenatal diagnosis of achondroplasia. *Gineco.eu.*12:56–58. DOI:10.18643/gieu.2016.56.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Mutations can be spontaneous or induced, somatic or germ-line, autosomal or sex-linked. Mutations can have many different effects on gene function depending on the type of nucleotide changes involved and the location of those mutations.
2. Spontaneous mutations occur in many ways, ranging from errors during DNA replication to damage caused to DNA bases as a result of normal cellular metabolism.
3. Mutations can be induced by many types of chemicals and radiation. These agents can damage both bases and the sugar-phosphate backbone of DNA molecules.
4. Single-gene mutations in humans cause a wide range of diseases. These mutations may be base-pair changes, insertions, deletions, and expanded DNA repeats.
5. Organisms counteract mutations using DNA repair systems including proofreading, mismatch repair, postreplication repair, photoreactivation repair, SOS repair, base excision repair, nucleotide excision repair, and double-strand break repair.
6. The Ames test allows scientists to estimate the mutagenicity and cancer-causing potential of chemical agents.
7. Transposable elements (TEs) can move within a genome, creating mutations and altering gene expression. In addition, TEs may contribute to evolution.

INSIGHTS AND SOLUTIONS

1. A rare dominant mutation expressed at birth was studied in humans. Records showed that six cases were discovered in 40,000 live births. Family histories revealed that in two cases, the mutation was already present in one of the parents. Calculate the spontaneous mutation rate for this mutation. What are some underlying assumptions that may affect our conclusions?

   Solution: Only four cases represent a new mutation. Because each live birth represents two gametes, the sample size is from 80,000 meiotic events. The rate is equal to

   $$4/80{,}000 = 1/20{,}000 = 5 \times 10^{-5}$$

   We have assumed that the mutant gene is fully penetrant and is expressed in each individual bearing it. If it is not fully penetrant, our calculation may be an underestimate because one or more mutations may have gone undetected. We have also assumed that the screening was 100 percent accurate. One or more mutant individuals may have been "missed," again leading to an underestimate. Finally, we assumed that the viabilities of the mutant and nonmutant individuals are equivalent and that they survive equally *in utero*. Therefore, our assumption is that the number of mutant individuals at birth is equal to the number at conception. If this were not true, our calculation would again be an underestimate.

2. Consider the following estimates:

   (a) There are 7×10^9 humans living on this planet.

   (b) Each individual has about 20,000 (0.2×10^5) genes.

   (c) The average mutation rate at each locus is 10^{-5}.

(continued)

Insights and Solutions—continued

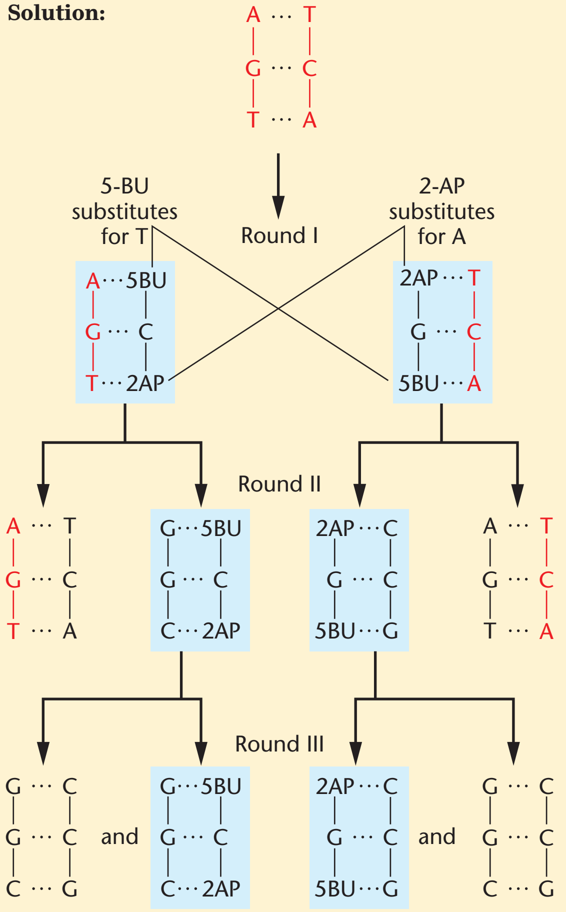
How many spontaneous mutations are currently present in the human population? Assuming that these mutations are equally distributed among all genes, how many new mutations have arisen in each gene in the human population?

Solution: First, since each individual is diploid, there are two copies of each gene per person, each arising from a separate gamete. Therefore, the total number of spontaneous mutations is

$$\begin{aligned}&(2 \times 0.2 \times 10^5 \text{ genes/individual})\\ &\quad \times (7 \times 10^9 \text{ individuals}) \times (10^{-5} \text{ mutations/gene})\\ &\quad = (0.4 \times 10^5) \times (7 \times 10^9) \times (10^{-5}) \text{ mutations}\\ &\quad = 2.8 \times 10^9 \text{ mutations in the population}\\ &2.8 \times 10^9 \text{ mutations}/0.2 \times 10^5 \text{ genes}\\ &\quad = 14 \times 10^4 \text{ mutations per gene in the population}\end{aligned}$$

3. The base analog 2-amino purine (2-AP) substitutes for adenine during DNA replication, but it may base-pair with cytosine. The base analog 5-bromouracil (5-BU) substitutes for thymine, but it may base-pair with guanine. Follow the double-stranded trinucleotide sequence shown at the top of the figure through three rounds of replication, assuming that, in the first round, both analogs are present and become incorporated wherever possible. Before the second and third round of replication, any unincorporated base analogs are removed. What final sequences occur?

Solution:



Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on how gene mutations arise and how cells repair DNA damage. At the same time, we found opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter,
(a) How do we know that mutations occur randomly?
(b) How do we know that certain chemicals and wavelengths of radiation induce mutations in DNA?
(c) How do we know that DNA repair mechanisms detect and correct the majority of spontaneous and induced mutations?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 340. These concepts relate to how gene mutations occur, their phenotypic effects, and how mutations can be repaired. Write a short essay contrasting how these concepts may differ between bacteria and eukaryotes.

3. What is a spontaneous mutation, and why are spontaneous mutations rare?

4. Why would a mutation in a somatic cell of a multicellular organism not necessarily result in a detectable phenotype?

5. Most mutations are thought to be deleterious. Why, then, is it reasonable to state that mutations are essential to the evolutionary process?

6. Why is a random mutation more likely to be deleterious than beneficial?

7. Most mutations in a diploid organism are recessive. Why?

8. What is the difference between a silent mutation and a neutral mutation?

9. Describe a tautomeric shift and how it may lead to a mutation.

10. Contrast and compare the mutagenic effects of deaminating agents, alkylating agents, and base analogs.

11. Why are frameshift mutations likely to be more detrimental than point mutations, in which a single pyrimidine or purine has been substituted?

12. Why are X rays more potent mutagens than UV radiation?

13. DNA damage brought on by a variety of natural and artificial agents elicits a wide variety of cellular responses involving numerous signaling pathways. In addition to the activation of DNA

repair mechanisms, there can be activation of pathways leading to apoptosis (programmed cell death) and cell-cycle arrest. Why would apoptosis and cell-cycle arrest often be part of a cellular response to DNA damage?

14. Contrast the various types of DNA repair mechanisms known to counteract the effects of UV radiation. What is the role of visible light in repairing UV-induced mutations?

15. Mammography is an accurate screening technique for the early detection of breast cancer in humans. Because this technique uses X rays diagnostically, it has been highly controversial. Can you explain why? What reasons justify the use of X rays for such a medical screening technique?

16. A significant number of mutations in the *HBB* gene that cause human β-thalassemia occur within introns or in upstream noncoding sequences. Explain why mutations in these regions often lead to severe disease, although they may not directly alter the coding regions of the gene.

17. Describe how the Ames test screens for potential environmental mutagens. Why is it thought that a compound that tests positively in the Ames test may also be carcinogenic?

18. What genetic defects result in the disorder xeroderma pigmentosum (XP) in humans? How do these defects create the phenotypes associated with the disorder?

19. Compare DNA transposons and retrotransposons. What properties do they share?

20. Speculate on how improved living conditions and medical care in the developed nations might affect human mutation rates, both neutral and deleterious.

21. In maize, a *Ds* or *Ac* transposon can alter the function of genes at or near the site of transposon insertion. It is possible for these elements to transpose away from their original insertion site, causing a reversion of the mutant phenotype. In some cases, however, even more severe phenotypes appear, due to events at or near the mutant allele. What might be happening to the transposon or the nearby gene to create more severe mutations?

22. It is estimated that about 0.2 percent of human mutations are due to TE insertions, and a much higher degree of mutational damage is known to occur in some other organisms. In what way might a TE insertion contribute positively to evolution?

23. In a bacterial culture in which all cells are unable to synthesize leucine (*leu*$^-$), a potent mutagen is added, and the cells are allowed to undergo one round of replication. At that point, samples are taken, a series of dilutions are made, and the cells are plated on either minimal medium or minimal medium containing leucine. The first culture condition (minimal medium) allows the growth of only *leu*$^+$ cells, while the second culture condition (minimal medium with leucine added) allows growth of all cells. The results of the experiment are as follows:

Culture Condition	Dilution	Colonies
Minimal medium	10^{-1}	18
Minimal medium + leucine	10^{-7}	6

What is the rate of mutation at the locus associated with leucine biosynthesis?

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

24. Presented here are hypothetical findings from studies of heterokaryons formed from seven human xeroderma pigmentosum cell strains:

	XP1	*XP2*	*XP3*	*XP4*	*XP5*	*XP6*	*XP7*
XP1	−						
XP2	−	−					
XP3	−	−	−				
XP4	+	+	+	−			
XP5	+	+	+	+	−		
XP6	+	+	+	+	−	−	
XP7	+	+	+	+	−	−	−

Note: + = complementation; − = no complementation

These data are measurements of the occurrence or nonoccurrence of unscheduled DNA synthesis in the fused heterokaryon. None of the strains alone shows any unscheduled DNA synthesis. Which strains fall into the same complementation groups? How many different groups are revealed based on these data? What can we conclude about the genetic basis of XP from these data?

25. Imagine yourself as one of the team of geneticists who launches a study of the genetic effects of high-energy radiation on the surviving Japanese population immediately following the atom bomb attacks at Hiroshima and Nagasaki in 1945. Demonstrate your insights into both chromosomal and gene mutation by outlining a short-term and long-term study that addresses these radiation effects. Be sure to include strategies for considering the effects on both somatic and germ-line tissues.

26. With the knowledge that radiation causes mutations, many assume that human-made forms of radiation are the major contributors to the mutational load in humans. What evidence suggests otherwise?

27. What evidence indicates that mutations in human DNA mismatch repair genes are related to certain forms of cancer?

28. Among Betazoids in the world of *Star Trek*®, the ability to read minds is under the control of a gene called *mindreader* (abbreviated *mr*). Most Betazoids can read minds, but rare recessive mutations in the *mr* gene result in two alternative phenotypes: *delayed-receivers* and *insensitives*. Delayed-receivers have some mind-reading ability but perform the task much more slowly than normal Betazoids. Insensitives cannot read minds at all. Betazoid genes do not have introns, so the gene only contains coding DNA. It is 3332 nucleotides in length, and Betazoids use a four-letter genetic code.

The following table shows some data from five unrelated *mr* mutations.

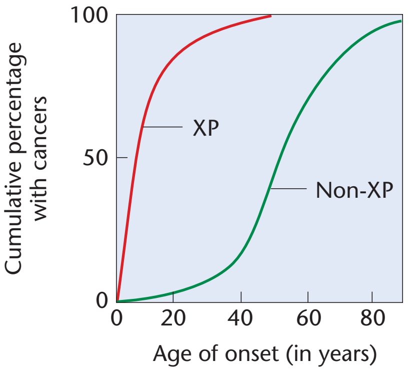
Mutation	Description of Mutation	Phenotype
mr-1	Nonsense mutation in codon 829	Delayed-receiver
mr-2	Missense mutation in codon 52	Delayed-receiver
mr-3	Deletion of nucleotides 83–150	Delayed-receiver
mr-4	Missense mutation in codon 192	Insensitive
mr-5	Deletion of nucleotides 83–93	Insensitive

For each mutation, provide a plausible explanation for why it gives rise to its associated phenotype and not to the other phenotype. For example, hypothesize why the *mr-1* nonsense mutation in codon 829 gives rise to the milder delayed-receiver phenotype rather than the more severe insensitive phenotype. Then repeat this type of analysis for the other mutations. (More than one explanation is possible, so be creative within *plausible* bounds!)

29. Skin cancer carries a lifetime risk nearly equal to that of all other cancers combined. Following is a graph [modified from K. H. Kraemer (1997). *Proc. Natl. Acad. Sci. (USA)* 94:11–14] depicting the age of onset of skin cancers in patients with or without XP, where the cumulative percentage of skin cancer is plotted against age. The non-XP curve is based on 29,757 cancers surveyed by the National Cancer Institute, and the curve representing those with XP is based on 63 skin cancers from the Xeroderma Pigmentosum Registry.

(a) Provide an overview of the information contained in the graph.

(b) Explain why individuals with XP show such an early age of onset.



30. It has been noted that most transposons in humans and other organisms are located in noncoding regions of the genome—regions such as introns, pseudogenes, and stretches of particular types of repetitive DNA. There are several ways to interpret this observation. Describe two possible interpretations. Which interpretation do you favor? Why?

31. Mutations in the *IL2RG* gene cause approximately 30 percent of severe combined immunodeficiency disorder (SCID) cases in humans. These mutations result in alterations to a protein component of cytokine receptors that are essential for proper development of the immune system. The *IL2RG* gene is composed of eight exons and contains upstream and downstream sequences that are necessary for proper transcription and translation. Below are some of the mutations observed. For each, explain its likely influence on the *IL2RG* gene product (assume its length to be 375 amino acids).

(a) Nonsense mutation in a coding region

(b) Insertion in Exon 1, causing frameshift

(c) Insertion in Exon 7, causing frameshift

(d) Missense mutation

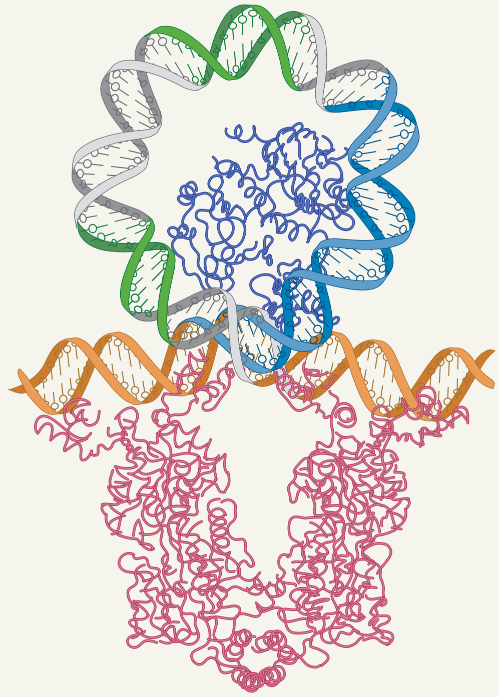
(e) Deletion in Exon 2, causing frameshift

(f) Deletion in Exon 2, in frame

(g) Large deletion covering Exons 2 and 3

16

Regulation of Gene Expression in Bacteria



A model showing how the *lac* repressor (red) and catabolite-activating protein (dark blue in center of DNA loop) bind to the *lac* operon promoter, creating a 93-base-pair repression loop in the *lac* regulatory DNA.

CHAPTER CONCEPTS

- In bacteria, regulation of gene expression is often linked to the metabolic needs of the cell.
- Efficient expression of genetic information in bacteria is dependent on intricate regulatory mechanisms that exert control over transcription.
- Mechanisms that regulate transcription are categorized as exerting either positive or negative control of gene expression.
- Bacterial genes that encode proteins with related functions tend to be organized in clusters and are often under coordinated control. Such clusters, including their adjacent regulatory sequences, are called operons.
- Transcription of genes within operons is either inducible or repressible.
- Often, the end product of a metabolic pathway induces or represses gene expression in that pathway.

Previous chapters have discussed how DNA is organized into genes, how genes store genetic information, and how this information is expressed through the processes of transcription and translation. We now consider one of the most fundamental questions in molecular genetics: *How is genetic expression regulated?* It is now clear that gene expression varies widely in bacteria under different environmental conditions. For example, detailed analysis of proteins in *Escherichia coli* shows that concentrations of the 4000 or so polypeptide chains encoded by the genome vary widely. Some proteins may be present in as few as 5 to 10 molecules per cell, whereas others, such as ribosomal proteins and the many proteins involved in the glycolytic pathway, are present in as many as 100,000 copies per cell. Although most bacterial gene products are present continuously at a basal level (a few copies), the concentration of these products can increase dramatically when required. Clearly, fundamental regulatory mechanisms must exist to control the expression of the genetic information.

In this chapter, we will explore regulation of gene expression in bacteria. As we have seen in several previous chapters, bacteria have been especially useful research organisms in genetics for a number of reasons. For one thing, they have extremely short reproductive cycles, and literally hundreds of generations, giving rise to billions of genetically identical bacteria, can be produced in overnight cultures. In addition, they can be studied in "pure culture," allowing mutant strains of genetically unique bacteria to be isolated and investigated separately.

Relevant to our current topic, bacteria also serve as an excellent model system for studies involving the induction of genetic transcription in response to changes in environmental conditions. While our primary focus will involve regulation of transcription, we will also consider instances where mRNAs contain sequences that serve as sensors that modulate their translation.

16.1 Bacteria Regulate Gene Expression in Response to Environmental Conditions

The regulation of gene expression has been extensively studied in bacteria, particularly in *E. coli*. Geneticists have learned that highly efficient genetic mechanisms have evolved in these organisms to turn transcription of specific genes on and off, depending on the cell's metabolic need for the respective gene products. Not only do bacteria respond to changes in their environment, but they also regulate gene activity associated with a variety of cellular activities, including the replication, recombination, and repair of their DNA, and with cell division.

The idea that microorganisms regulate the synthesis of their gene products is not a new one. As early as 1900, it was shown that when lactose (a galactose and glucose-containing disaccharide) is present in the growth medium of yeast, the organisms synthesize enzymes required for lactose metabolism. When lactose is absent, synthesis diminishes to a basal level. Soon thereafter, investigators generalized that bacteria adapt to their environment, producing certain enzymes only when specific chemical substrates are present. These are now referred to as **inducible enzymes.** In contrast, enzymes that are produced continuously, regardless of the chemical makeup of the environment, are called **constitutive enzymes**.

More recent investigation has revealed a contrasting system, whereby the presence of a specific molecule *inhibits* gene expression. Such molecules are usually end products of anabolic biosynthetic pathways. For example, utilizing a multistep metabolic pathway, the amino acid tryptophan can be synthesized by bacterial cells. If a sufficient supply of tryptophan is present in the environment or culture medium, then it is inefficient for the organism to expend energy to synthesize the enzymes necessary for tryptophan production. A mechanism has therefore evolved whereby tryptophan plays a role in repressing the transcription of mRNA needed for producing tryptophan-synthesizing enzymes. In contrast to the inducible system controlling lactose metabolism, the system governing tryptophan expression is said to be **repressible**.

Regulation, whether of the inducible or repressible type, may be under either **negative** or **positive control.** Under negative control, genetic expression occurs *unless it is shut off by some form of a regulator molecule*. In contrast, under positive control, transcription occurs *only if a regulator molecule directly stimulates RNA production*. In theory, either type of control or a combination of the two can govern inducible or repressible systems. Our discussion in the ensuing sections of this chapter will help clarify these contrasting systems of regulation. The enzymes involved in lactose digestion and tryptophan synthesis are under negative control.

16.2 Lactose Metabolism in *E. coli* Is Regulated by an Inducible System

Beginning in 1946 with the studies of Jacques Monod and continuing through the next decade with significant contributions by Joshua Lederberg, François Jacob, and André Lwoff, genetic and biochemical evidence concerning lactose metabolism was amassed. Research provided insights into the way in which the gene activity is repressed when lactose is absent but induced when it is available. In the presence of lactose, the concentration of the enzymes responsible for its metabolism increases rapidly from a few molecules to thousands per cell. The enzymes responsible for lactose metabolism are thus inducible, and lactose serves as the inducer.

In bacteria, genes that code for enzymes with related functions (for example, the set of genes involved with lactose metabolism) tend to be organized in clusters on the bacterial chromosome, and transcription of these genes is often under the coordinated control of a single regulatory region. Such clusters, including their adjacent regulatory sequences, are called **operons.** The location of the regulatory region is almost always upstream (5′) of the gene cluster it controls. Because the regulatory region is on the same strand as those genes, we refer to it as a ***cis*-acting site.** *Cis*-acting regulatory regions bind molecules that control transcription of the gene cluster. Such molecules are called ***trans*-acting factors.** Events at the regulatory site determine whether the genes are transcribed into mRNA and thus whether the corresponding enzymes or other protein products may be synthesized from the genetic information in the mRNA. Binding of a *trans*-acting element at a *cis*-acting site can regulate the gene cluster either negatively (by turning off transcription) or positively (by turning on transcription of genes in the cluster). In this section, we discuss how transcription of such bacterial gene clusters is coordinately regulated.

The discovery of a regulatory gene and a regulatory site that are part of the gene cluster was paramount to the understanding of how gene expression is controlled in the system. Neither of these regulatory elements encodes enzymes necessary for lactose metabolism—the function of the three genes in the cluster. As illustrated in **Figure 16.1**, the three structural genes and the adjacent regulatory site constitute the **lactose (*lac*) operon.** Together, the entire gene cluster functions in an integrated fashion to provide a rapid response to the presence or absence of lactose.

Structural Genes

Genes coding for the primary structure of an enzyme are called **structural genes.** There are three structural genes in the *lac* operon. The *lacZ* gene encodes **β-galactosidase,** an enzyme whose primary role is to convert the disaccharide lactose to the monosaccharides glucose and galactose (**Figure 16.2**). This conversion is essential if lactose is to serve as the primary

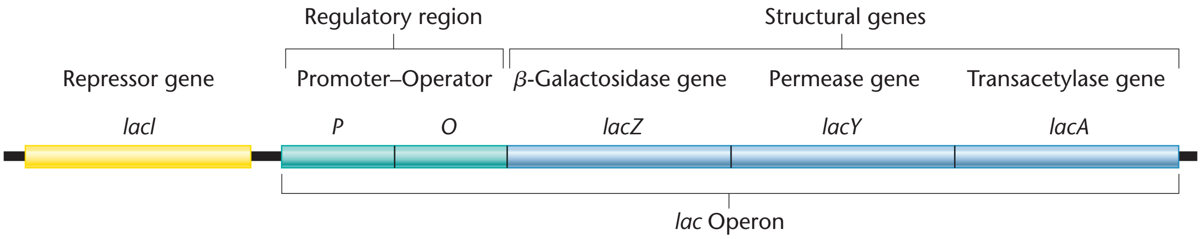


FIGURE 16.1 A simplified overview of the genes and regulatory units involved in the control of lactose metabolism. (The regions within this stretch of DNA are not drawn to scale.) A more detailed model will be developed later in this chapter. (See Figure 16.10.)

energy source in glycolysis. The second gene, *lacY,* specifies the primary structure of **permease,** an enzyme that facilitates the entry of lactose into the bacterial cell. The third gene, *lacA,* codes for the enzyme **transacetylase.** While its physiological role is still not completely clear, it may be involved in the removal of toxic by-products of lactose digestion from the cell.

To study the genes coding for these three enzymes, researchers isolated numerous mutations that lacked the function of one or the other enzyme. Such *lac*$^-$ mutants were first isolated and studied by Joshua Lederberg. Mutant cells that fail to produce active β-galactosidase (*lacZ*$^-$) or permease (*lacY*$^-$) are unable to use lactose as an energy source. Mutations were also found in the transacetylase gene. Mapping studies by Lederberg established that all three genes are closely linked or contiguous to one another on the bacterial chromosome, in the order *Z*–*Y*–*A* (see Figure 16.1).

Knowledge of their close linkage led to another discovery relevant to what later became known about the regulation of structural genes: All three genes are transcribed as a single unit, resulting in a so-called *polycistronic mRNA* (Figure 16.3) (recall that *cistron* refers to the part of a nucleotide sequence coding for a single gene). This results in the coordinate regulation of all three genes, since a single-message RNA is simultaneously translated into all three gene products.

The Discovery of Regulatory Mutations

How does lactose stimulate transcription of the *lac* operon and induce the synthesis of the enzymes for which it codes? A partial answer came from studies using **gratuitous inducers,** chemical analogs of lactose such as the sulfur-containing analog **isopropylthiogalactoside (IPTG),** shown in Figure 16.4. Gratuitous inducers behave like natural inducers, but they do not serve as substrates for the enzymes that are subsequently synthesized. Their discovery provides strong evidence that the primary induction event does *not* depend on the interaction between the inducer and the enzyme.

What, then, is the role of lactose in induction? The answer to this question requires the study of another

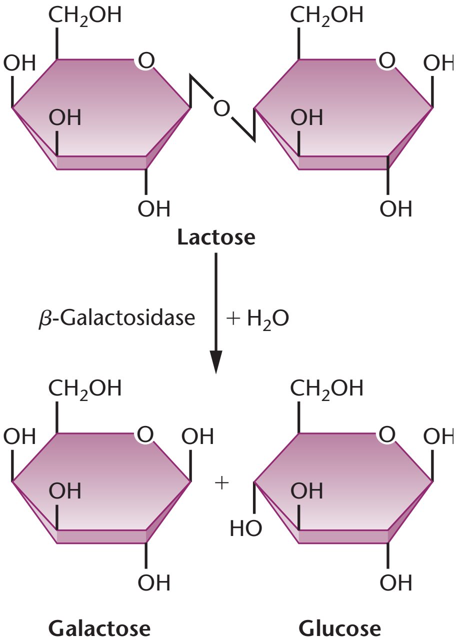


FIGURE 16.2 The catabolic conversion of the disaccharide lactose into its monosaccharide units, galactose and glucose.

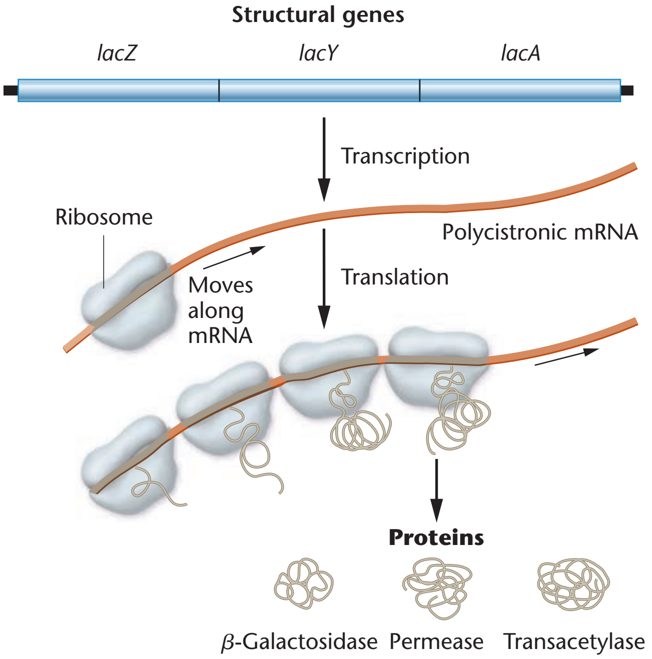


FIGURE 16.3 The structural genes of the *lac* operon are transcribed into a single polycistronic mRNA, which is translated simultaneously by several ribosomes into the three enzymes encoded by the operon.

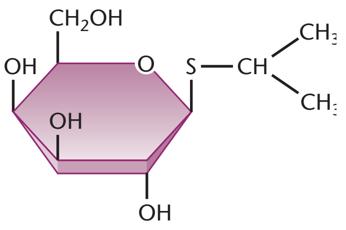
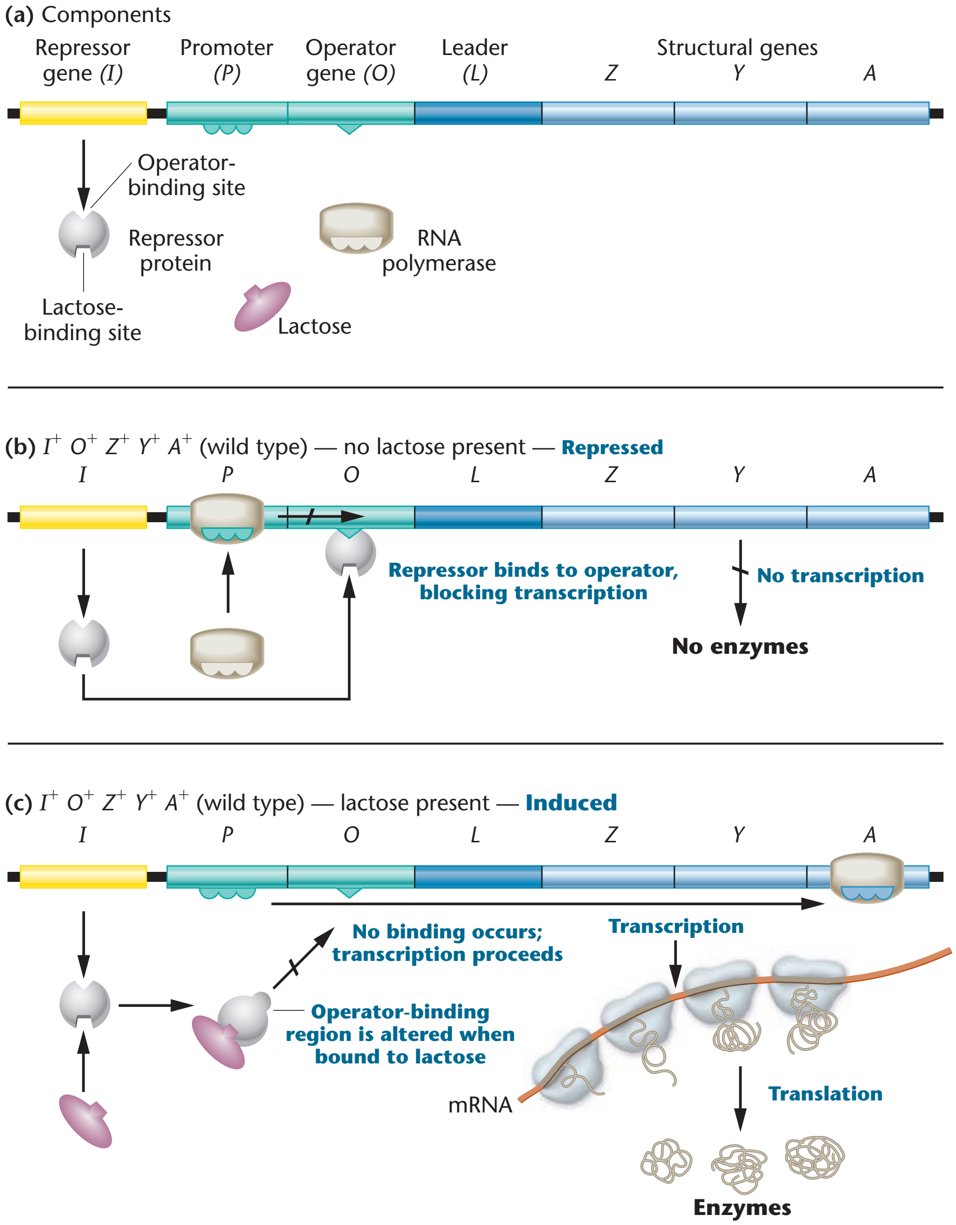


FIGURE 16.4 The gratuitous inducer isopropylthiogalactoside (IPTG).

class of mutations described as **constitutive mutations.** In cells bearing these types of mutations, enzymes are produced regardless of the presence or absence of lactose. Studies of the constitutive mutation *lacI*$^-$ mapped the mutation to a site on the bacterial chromosome close to, but not part of, the structural genes *lacZ, lacY,* and *lacA*. This mutation led researchers to discover the *lacI* gene, which is appropriately called a **repressor gene.** A second set of constitutive mutations producing effects identical to those of *lacI*$^-$ is present in a region immediately adjacent to the structural genes. This class of mutations, designated *lac O*^{C}, is located in the **operator region** of the operon. In both types of constitutive mutants, the enzymes are produced continually, inducibility is eliminated, and gene regulation has been lost.

The Operon Model: Negative Control

Around 1960, Jacob and Monod proposed a hypothetical mechanism involving negative control that they called the **operon model,** in which a group of genes is regulated


(a) Components

(b) I^+ O^+ Z^+ Y^+ A^+ (wild type) — no lactose present — **Repressed**

(c) I^+ O^+ Z^+ Y^+ A^+ (wild type) — lactose present — **Induced**

I
P
O
L
Z
Y
A
No binding occurs; transcription proceeds
Transcription
Operator-binding region is altered when bound to lactose
mRNA
Translation
Enzymes

FIGURE 16.5 The components of the wild-type *lac* operon and the response in the absence and presence of lactose.

and expressed together as a unit. As we saw in Figure 16.1, the *lac* operon they proposed consists of the *Z, Y,* and *A* structural genes, as well as the adjacent sequences of DNA referred to as the *operator region*. They argued that the *lacI* gene regulates the transcription of the structural genes by producing a **repressor molecule** and that the repressor is **allosteric,** meaning that the molecule reversibly interacts with another molecule, undergoing both a conformational change in three-dimensional shape and a change in chemical activity. Figure 16.5 illustrates the components of the *lac* operon as well as the action of the *lac* repressor in the presence and absence of lactose.

Jacob and Monod suggested that the repressor normally binds to the DNA sequence of the operator region. When it does so, it inhibits the action of RNA polymerase, effectively repressing the transcription of the structural genes [Figure 16.5(b)]. However, when lactose is present, this sugar binds to the repressor and causes an allosteric (conformational) change. The change alters the binding site of the repressor, rendering it incapable of interacting with operator DNA [Figure 16.5(c)]. In the absence of the repressor–operator interaction, RNA polymerase transcribes the structural genes, and the enzymes necessary for lactose metabolism are produced. Because transcription occurs only when the repressor *fails* to bind to the operator region, regulation is said to be under *negative control*.

To summarize, the operon model invokes a series of molecular interactions between proteins, inducers, and DNA to explain the efficient regulation of structural gene expression. In the absence of lactose, the enzymes encoded by the genes are not needed and the expression of genes encoding these enzymes is repressed. When lactose is present, it indirectly induces the activation of the genes by binding with the repressor.* If all lactose is metabolized, none is available to bind to the repressor, which is again free to bind to operator DNA and to repress transcription.

Both the I^- and O^C constitutive mutations interfere with these molecular interactions, allowing continuous transcription of the structural genes. In the case of the I^- mutant, seen in Figure 16.6(a), the repressor protein is altered or absent and cannot bind to the operator region,

*Technically, the inducer is allolactose, an isomer of lactose. When lactose enters the bacterial cell, some of it is converted to allolactose by the β-galactosidase enzyme.

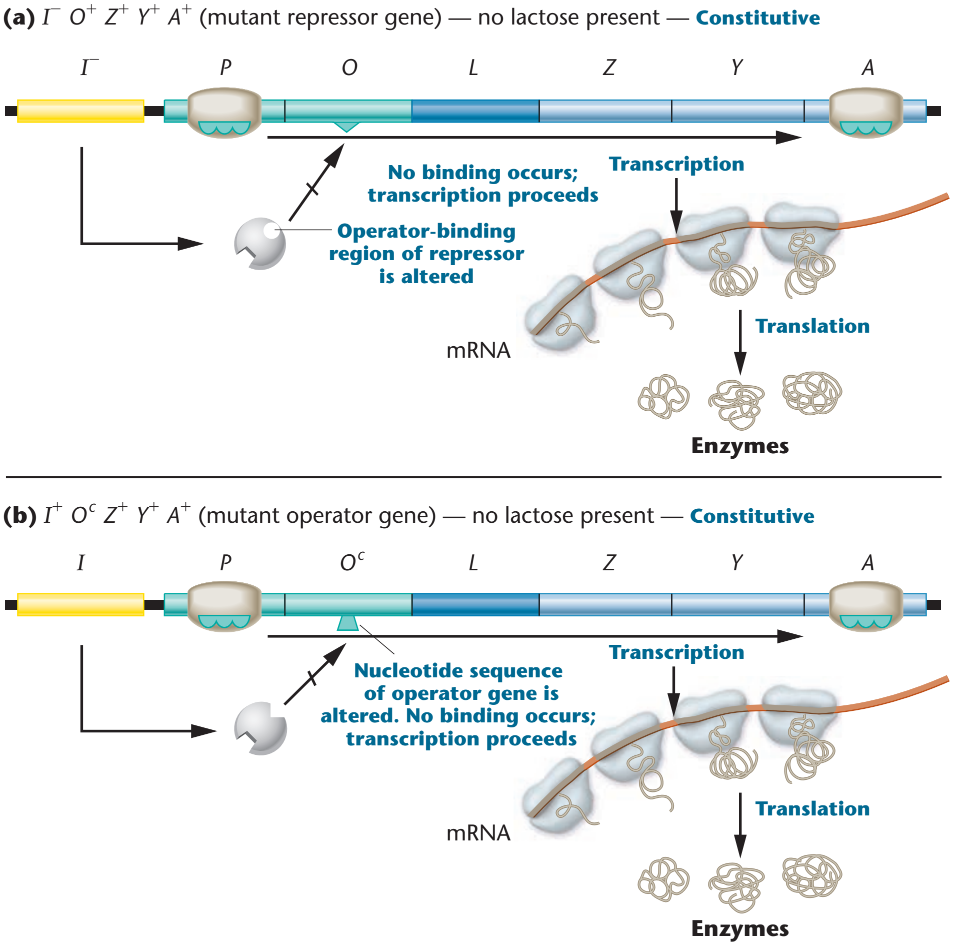

(a) I^- O^+ Z^+ Y^+ A^+ (mutant repressor gene) — no lactose present — Constitutive
I^-
P
O
L
Z
Y
A
No binding occurs; transcription proceeds
Transcription
Operator-binding region of repressor is altered
mRNA
Translation
Enzymes
(b) I^+ O^c Z^+ Y^+ A^+ (mutant operator gene) — no lactose present — Constitutive
I
P
O^c
L
Z
Y
A
Transcription
Nucleotide sequence of operator gene is altered. No binding occurs; transcription proceeds
mRNA
Translation
Enzymes

FIGURE 16.6 The response of the *lac* operon in the absence of lactose when a cell bears either the I^- or the O^C mutation.

so the structural genes are always turned on. In the case of the O^C mutant [Figure 16.6(b)], the nucleotide sequence of the operator DNA is altered and will not bind with a normal repressor molecule. The result is the same: The structural genes are always transcribed.

Genetic Proof of the Operon Model

The operon model is a good one because it leads to three major predictions that can be tested to determine its validity. The major predictions to be tested are that (1) the *I* gene produces a diffusible product (that is, a *trans*-acting product), (2) the *O* region is involved in regulation but does not produce a product (it is *cis*-acting), and (3) the *O* region must be adjacent to the structural genes to regulate transcription.

The creation of partially diploid bacteria allows us to assess these assumptions, particularly those that predict the presence of *trans*-acting regulatory elements. For example, the F plasmid may contain chromosomal genes (Chapter 6), in which case it is designated F′. When an F^- cell acquires such a plasmid, it contains its own chromosome plus one or more additional genes present in the plasmid. This host cell is thus a **merozygote,** a cell that is diploid for certain added genes (but not for the rest of the chromosome). The use of such a plasmid makes it possible, for example, to introduce an I^+ gene into a host cell whose genotype is I^-, or to introduce an O^+ region into a host cell of genotype O^C. The Jacob–Monod operon model predicts how regulation should be affected in such cells. Adding an I^+ gene to an I^- cell should restore inducibility, because the normal wild-type repressor, which is a *trans*-acting factor, would be produced by the inserted I^+ gene. In contrast, adding an O^+ region to an O^C cell should have no effect on constitutive enzyme production, since regulation depends on an O^+ region being located immediately adjacent to the structural genes—that is, O^+ is a *cis*-acting element.

Results of these experiments are shown in Table 16.1, where *Z* represents the structural genes (and the inserted genes are listed after the designation F′). In both cases described above, the Jacob–Monod model is upheld (part B of Table 16.1). Part C of the table shows the reverse experiments, where either an I^- gene or an O^C region is added to cells of normal inducible genotypes. As the model predicts, inducibility is maintained in these partial diploids.

Another prediction of the operon model is that certain mutations in the *I* gene should have the opposite effect of I^-. That is, instead of being constitutive because the repressor cannot bind the operator, mutant repressor molecules should be produced that cannot interact with the inducer, lactose. Thus, these repressors would always bind to the operator sequence, and the structural genes would be permanently repressed. In cases like this, the presence of an additional I^+ gene would have little or no effect on repression.

TABLE 16.1 A Comparison of Gene Activity (+ or −) in the Presence or Absence of Lactose for Various *E. coli* Genotypes

	Presence of β-Galactosidase Activity	
Genotype	Lactose Present	Lactose Absent
A. $I^+O^+Z^+$	+	−
$I^+O^+Z^-$	−	−
$I^-O^+Z^+$	+	+
$I^+O^CZ^+$	+	+
B. $I^-O^+Z^+/F'I^+$	+	−
$I^+O^CZ^+/F'O^+$	+	+
C. $I^+O^+Z^+/F'I^-$	+	−
$I^+O^+Z^+/F'O^C$	+	−
D. $I^SO^+Z^+$	−	−
$I^SO^+Z^+/F'I^+$	−	−

*Note: In parts B to D, most genotypes are partially diploid, containing an F factor plus attached genes (F′).

In fact, such a mutation, I^S, was discovered wherein the operon, as predicted, is "superrepressed," as shown in part D of Table 16.1 (and depicted in Figure 16.7). An additional I^+ gene does not effectively relieve repression of gene activity. These observations are consistent with the idea that the repressor contains separate DNA-binding domains and inducer-binding domains.

Isolation of the Repressor

Although Jacob and Monod's operon theory succeeded in explaining many aspects of genetic regulation in bacteria, the nature of the repressor molecule was not known when their landmark paper was published in 1961. While they had assumed that the allosteric repressor was a protein, RNA was also a viable candidate because the activity of the molecule required the ability to bind to DNA. Despite many attempts to isolate and characterize the hypothetical repressor molecule, no direct chemical evidence was forthcoming. A single *E. coli* cell contains no more than ten or so copies of the *lac* repressor, and direct chemical identification of ten molecules in a population of millions of proteins and RNAs in a single cell presented a tremendous challenge.

In 1966, Walter Gilbert and Benno Müller-Hill reported the isolation of the *lac* repressor in partially purified form. To facilitate the isolation, they used a *regulator quantity* (I^q) mutant strain that contains about ten times as much repressor as do wild-type *E. coli* cells. Also instrumental in their success was the use of the gratuitous inducer IPTG, which binds to the repressor, and the technique of **equilibrium dialysis.** In this technique, extracts of I^q cells were placed

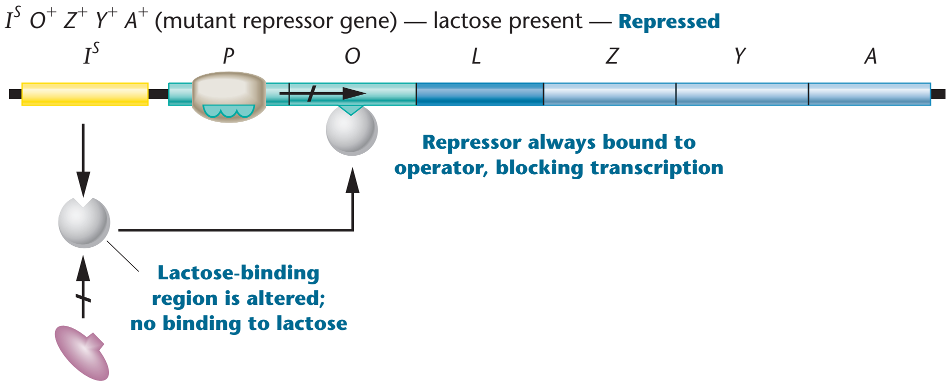

I^S O^+ Z^+ Y^+ A^+ (mutant repressor gene) — lactose present — Repressed
I^S
P
O
L
Z
Y
A
Repressor always bound to operator, blocking transcription
Lactose-binding region is altered; no binding to lactose

FIGURE 16.7 The response of the *lac* operon in the presence of lactose in a cell bearing the I^S mutation.

in a dialysis bag and allowed to attain equilibrium with an external solution of radioactive IPTG, a molecule small enough to diffuse freely in and out of the bag. At equilibrium, the concentration of radioactive IPTG was higher inside the bag than in the external solution, indicating that an IPTG-binding material was present in the cell extract and was too large to diffuse across the wall of the bag.

NOW SOLVE THIS

16.1 Even though the *lac Z, Y,* and *A* structural genes are transcribed as a single polycistronic mRNA, each gene contains the initiation and termination signals essential for translation. Predict what will happen when a cell growing in the presence of lactose contains a deletion of one nucleotide (a) early in the *Z* gene and (b) early in the *A* gene.

- **HINT:** *This problem requires you to combine your understanding of the genetic expression of the* lac *operon with that of the genetic code, frameshift mutations, and termination of transcription. The key to its solution is to consider the effect of the loss of one nucleotide within a polycistronic mRNA.*

For more practice, see Problems 5–7.

Ultimately, the IPTG-binding material was purified and shown to have various characteristics of a protein. In contrast, extracts of I^- constitutive cells having no *lac* repressor *activity* did not exhibit IPTG binding, strongly suggesting that the isolated protein was the repressor molecule.

To confirm this thinking, Gilbert and Müller-Hill grew *E. coli* cells in a medium containing radioactive sulfur and then isolated the IPTG-binding protein, which was labeled in its sulfur-containing amino acids. Next, this protein was mixed with DNA from a strain of phage lambda (λ) carrying the *lacO*$^+$ gene. When the two substances are separate, the DNA sediments at 40*S* and the IPTG-binding protein sediments at 7*S*. However, when the DNA and protein were mixed and sedimented in a gradient, using ultracentrifugation, the radioactive protein sedimented at the same rate as did the DNA, indicating that the protein binds to the DNA. Further experiments showed that the IPTG-binding, or repressor, protein binds only to DNA containing the *lac* region and does not bind to *lac* DNA containing an operator-constitutive O^C mutation.

16.3 The Catabolite-Activating Protein (CAP) Exerts Positive Control over the *lac* Operon

As described in the preceding discussion of the *lac* operon, the role of β-galactosidase is to cleave lactose into its components, glucose and galactose. Then, to be used by the cell, the galactose, too, must be converted to glucose. What if the cell found itself in an environment that contained ample amounts of both lactose *and* glucose? Given that glucose is the preferred carbon source for *E. coli,* it would not be energetically efficient for a cell to induce transcription of the *lac* operon, since what it really needs—glucose—is already present. As we will see next, still another molecular component, called the **catabolite-activating protein (CAP),** is involved in diminishing the expression of the *lac* operon when glucose is present. This inhibition is called **catabolite repression**.

To understand CAP and its role in regulation, let's backtrack for a moment to review the system depicted in Figure 16.5. When the *lac* repressor is bound to the inducer (lactose), the *lac* operon is activated, and RNA polymerase transcribes the structural genes. As stated earlier in the text (see Chapter 14), transcription is initiated as a result of the binding that occurs between RNA polymerase and

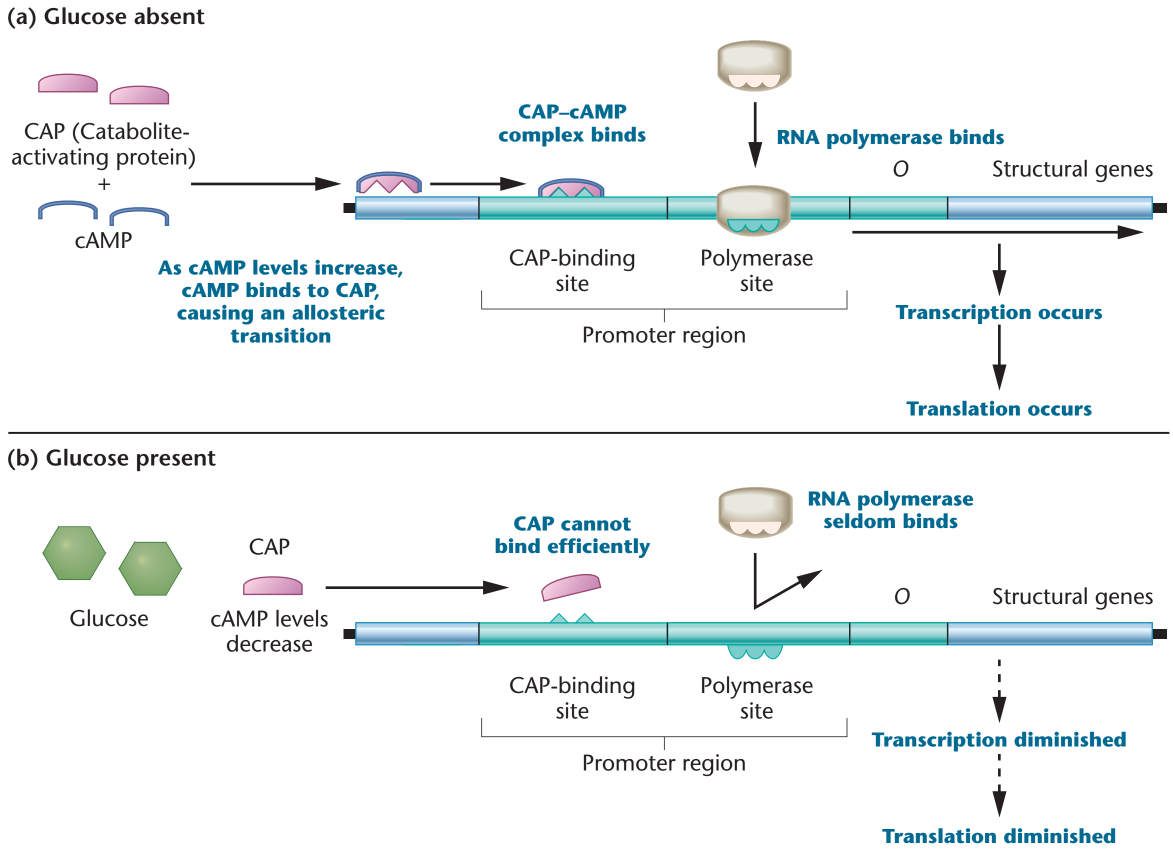


FIGURE 16.8 Catabolite repression. (a) In the absence of glucose, cAMP levels increase, resulting in the formation of a cAMP–CAP complex, which binds to the CAP site of the promoter, stimulating transcription. (b) In the presence of glucose, cAMP levels decrease, cAMP–CAP complexes are not formed, and transcription is not stimulated.

the nucleotide sequence of the **promoter region,** found upstream (5′) from the initial coding sequences. Within the *lac* operon, the promoter is found between the *I* gene and the operator region (*O*) (see Figure 16.1). Careful examination has revealed that polymerase binding is never very efficient unless CAP is also present to facilitate the process.

The mechanism is summarized in **Figure 16.8**. In the absence of glucose and under inducible conditions, CAP exerts positive control by binding to the CAP site, facilitating RNA-polymerase binding at the promoter, and thus transcription. Therefore, for maximal transcription of the structural genes to occur, the repressor must be bound by lactose (so as not to repress operon expression) *and* CAP must be bound to the CAP-binding site.

This leads to the central question about CAP: How does the presence of glucose inhibit CAP binding? The answer involves still another molecule, **cyclic adenosine monophosphate (cAMP),** upon which CAP binding is dependent. *In order to bind to the promoter, CAP must be bound to cAMP*. The level of cAMP is itself dependent on an enzyme, **adenyl cyclase,** which catalyzes the conversion of ATP to cAMP (see **Figure 16.9**).* The role of glucose in catabolite repression is now clear. It inhibits the activity of adenyl cyclase, causing a decline in the level of cAMP in the cell. Under this condition, CAP cannot form the cAMP–CAP complex essential to the positive control of transcription of the *lac* operon.

Like the *lac* repressor, CAP and cAMP–CAP binding have been examined by X-ray crystallography. CAP is a dimer that inserts into adjacent regions of a specific nucleotide sequence of the DNA making up the *lac* promoter. The cAMP–CAP complex, when bound to DNA, bends it, causing it to assume a new conformation.

Binding studies in solution further clarify the mechanism of gene activation. Alone, neither cAMP–CAP nor RNA polymerase has a strong tendency to bind to *lac* promoter DNA, nor does either molecule have a strong affinity for the other. However, when both are together in the presence of the *lac* promoter DNA, a tightly bound complex is formed, an example of what is called **cooperative binding.** In the case of cAMP–CAP and the *lac* operon, this phenomenon illustrates the high degree of specificity

*Because of its involvement with cAMP, CAP is also called *cyclic AMP receptor protein (CRP)*, and the gene encoding the protein is named *crp*. Since the protein was first named CAP, we will adhere to the initial nomenclature.

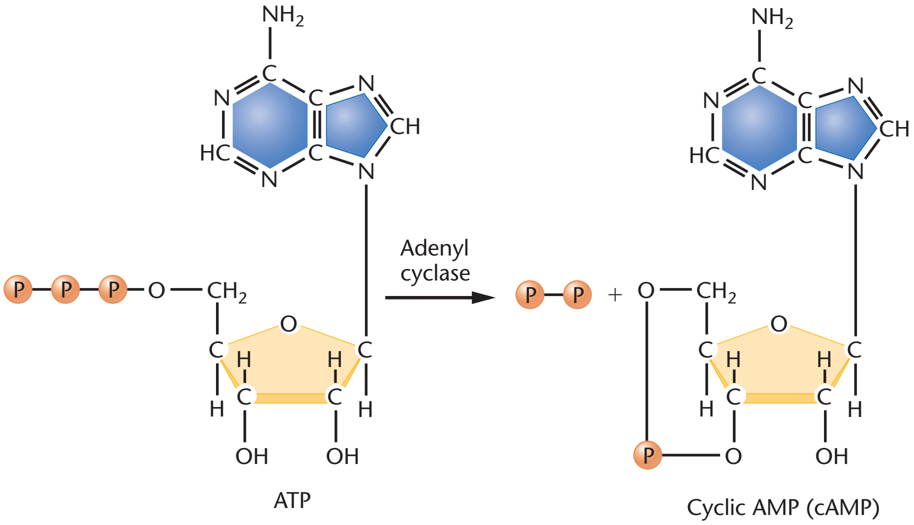


FIGURE 16.9 The formation of cAMP from ATP, catalyzed by adenyl cyclase.

that is involved in the genetic regulation of just one small group of genes.

NOW SOLVE THIS

16.2 Predict the level of genetic activity of the *lac* operon as well as the status of the *lac* repressor and the CAP protein under the cellular conditions listed in the accompanying table.

	Lactose	Glucose
(a)	−	−
(b)	+	−
(c)	−	+
(d)	+	+

■ **HINT:** *This problem asks you to combine your knowledge of the Jacob–Monod model of the regulation of the* lac *operon with your understanding of how catabolite repression impacts on this model. The key to its solution is to keep in mind that regulation involving lactose is a negative control system, while regulation involving glucose and catabolite repression is a positive control system.*

Regulation of the *lac* operon by catabolite repression results in efficient energy use, because the presence of glucose will override the need for the metabolism of lactose, if lactose is also available to the cell. In contrast to the negative regulation conferred by the *lac* repressor, the action of cAMP–CAP constitutes positive regulation. Thus, a combination of positive and negative regulatory mechanisms determines transcription levels of the *lac* operon. Catabolite repression involving CAP has also been observed for other inducible operons, including those controlling the metabolism of galactose and arabinose.

16.4 Crystal Structure Analysis of Repressor Complexes Has Confirmed the Operon Model

We now have a thorough knowledge of the biochemical nature of the regulatory region of the *lac* operon, including the precise locations of its various components relative to one another (Figure 16.10). In 1996, Mitchell Lewis, Ponzy Lu, and their colleagues succeeded in determining the crystal structure of the *lac* repressor, as well as the structure of the repressor bound to the inducer and to operator DNA. As a result, previous information that was based on genetic and biochemical data has now been complemented with the missing structural interpretation. Together, these contributions provide a nearly complete picture of the regulation of the operon.

The repressor, as the gene product of the *I* gene, is a monomer consisting of 360 amino acids. Within this monomer, the region of inducer binding has been identified [Figure 16.11(a)]. Although dimers [Figure 16.11(b)] can also bind the inducer, the functional repressor is a homotetramer (that is, it contains four copies of the monomer). The tetramer can be cleaved with a protease under controlled conditions to yield five fragments. Four are derived from the N-terminal ends of the tetramer subunits, and they bind to operator DNA. The fifth fragment is the remaining core of the tetramer, derived from the COOH-terminal ends; it binds to lactose and gratuitous inducers such as IPTG. Analysis has revealed that each tetramer can bind to two symmetrical operator DNA helices at a time.

The operator DNA that was previously defined by mutational studies (*lacO*^{C}) and confirmed by DNA-sequencing

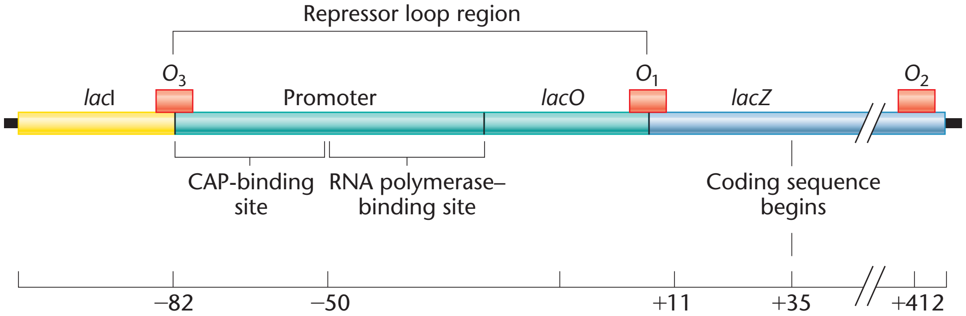


FIGURE 16.10 The various regulatory regions involved in the control of genetic expression of the *lac* operon, as described in the text. The numbers on the bottom scale represent nucleotide sites upstream and downstream from the initiation of transcription.

analysis is located just upstream from the beginning of the actual coding sequence of the *lacZ* gene. Crystallographic studies show that the actual region of repressor binding of this primary operator, O_1, consists of 21 base pairs. Two other auxiliary operator regions have been identified, as shown in Figure 16.10. One, O_2, is 401 base pairs downstream from the primary operator, within the *lacZ* gene. The other, O_3, is 93 base pairs upstream from O_1, just beyond the CAP site. *In vivo*, all three operators must be bound for maximum repression.

Binding by the repressor simultaneously at two operator sites distorts the conformation of DNA, causing it to bend away from the repressor. When a model is created to demonstrate dual binding of operators O_1 and O_3 [Figure 16.11(c)], the 93 base pairs of DNA that intervene must jut out, forming what is called a **repression loop.**

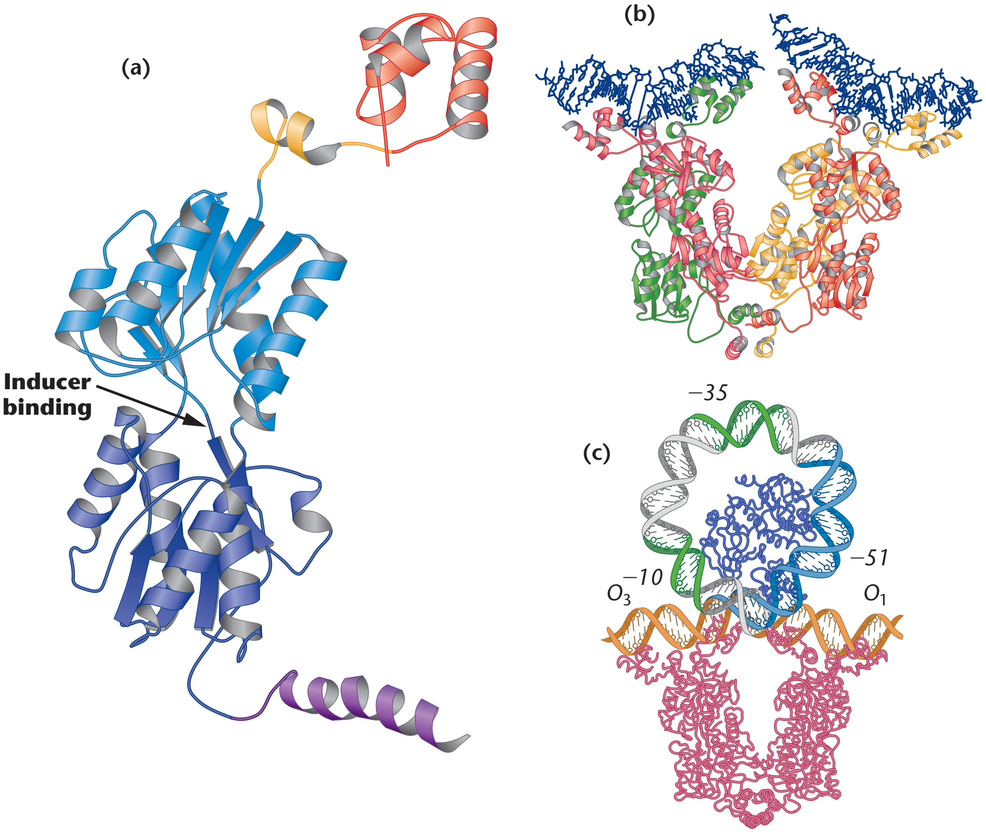


FIGURE 16.11 Models of the *lac* repressor and its binding to operator sites in DNA, as generated from crystal structure analysis. (a) The repressor monomer, showing the inducer-binding site. The DNA-binding region is shown in red. (b) The repressor tetramer bound to two 21-base-pair segments of operator DNA (shown in dark blue). (c) The repressor (shown in pink) and CAP (shown in dark blue) bound to the *lac* DNA. Binding to operator regions O_1 and O_3 creates a 93-base-pair repression loop of promoter DNA.

This model positions the promoter region that binds RNA polymerase on the inside of the loop, which prevents access by the polymerase during repression. In addition, the repression loop positions the CAP-binding site in a way that facilitates CAP interaction with RNA polymerase upon subsequent induction. The DNA looping caused by repression in this model is similar to configuration changes that are predicted to occur in eukaryotic systems (see Chapter 17).

Studies have also defined the three-dimensional conformational changes that accompany the interactions with the inducer molecules. Taken together, the crystallographic studies have brought a new level of understanding of the regulatory process occurring within the *lac* operon, confirming the findings and predictions of Jacob and Monod in their model set forth over 40 years ago, which was based strictly on genetic observations.

16.5 The Tryptophan (*trp*) Operon in *E. coli* Is a Repressible Gene System

Although the process of induction had been known for some time, it was not until 1953 that Monod and colleagues discovered a repressible operon. When grown in minimal medium (see Chapter 6), wild-type *E. coli* produce the enzymes necessary for the biosynthesis of amino acids as well as many other essential macromolecules. Focusing his studies on the amino acid tryptophan and the enzyme **tryptophan synthetase,** Monod discovered that if tryptophan is present in sufficient quantity in the growth medium, the enzymes necessary for its synthesis are not produced. It is energetically advantageous for bacteria to repress expression of genes involved in tryptophan synthesis when ample tryptophan is present in the growth medium.

Further investigation showed that a series of enzymes encoded by five contiguous genes on the *E. coli* chromosome are involved in tryptophan synthesis. These genes are part of an operon, and in the presence of tryptophan, all are coordinately repressed and none of the enzymes are produced. Because of the great similarity between this repression and the induction of enzymes for lactose metabolism, Jacob and Monod proposed a model of gene regulation analogous to the *lac* system (the updated version is shown in **Figure 16.12**).

To account for repression, Jacob and Monod suggested the presence of a *normally inactive repressor* that alone cannot interact with the operator region of the operon. However, the repressor is an allosteric molecule that can bind to tryptophan. When tryptophan is present, the resultant complex of repressor and tryptophan attains a new conformation that binds to the operator, repressing transcription. Thus, when tryptophan, the end product of this anabolic pathway, is present, the system is repressed and enzymes are not made. Since the regulatory complex inhibits transcription of the operon, this repressible system is under negative control. And as tryptophan participates in repression, it is referred to as a **corepressor** in this regulatory scheme.

Evidence for the *trp* Operon

Support for the concept of a repressible operon was soon forthcoming, based primarily on the isolation of two distinct categories of constitutive mutations. The first class, *trpR*$^-$, maps at a considerable distance from the structural genes. This locus represents the gene coding for the repressor. Presumably, the mutation inhibits either the repressor's interaction with tryptophan or repressor formation entirely. Whichever the case, repression never occurs in cells with the *trpR*$^-$ mutation. As expected, if the *trpR*$^+$ gene encodes a functional repressor molecule, the presence of a copy of this gene will restore repressibility.

The second constitutive mutation is analogous to that of the operator of the lactose operon, because it maps immediately adjacent to the structural genes. Furthermore, the insertion of a plasmid bearing a wild-type operator gene into mutant cells does not restore repression. This is what would be predicted if the mutant operator no longer interacts with the repressor–tryptophan complex.

The entire *trp* operon has now been well defined, as shown in Figure 16.12. The five contiguous structural genes (*trpE*, *D*, *C*, *B*, and *A*) are transcribed as a polycistronic message directing translation of the enzymes that catalyze the biosynthesis of tryptophan. As in the *lac* operon, a promoter region (*trpP*) represents the binding site for RNA polymerase, and an operator region (*trpO*) binds the repressor. In the absence of binding, transcription is initiated within the *trpP*–*trpO* region and proceeds along a **leader sequence** 162 nucleotides prior to the first structural gene (*trpE*). Within that leader sequence, still another regulatory site has been demonstrated, called an *attenuator*—the subject of Section 16.6. As we will see, this regulatory unit is an integral part of this operon's control mechanism.

EVOLVING CONCEPT OF THE GENE

The groundbreaking work of Jacob, Monod, and Lwoff in the early 1960s, which established the operon model for the regulation of gene expression in bacteria, expanded the concept of the gene to include noncoding regulatory sequences that are present upstream (5′) from the coding region. In bacterial operons, the transcription of several contiguous structural genes whose products are involved in the same biochemical pathway is regulated in a coordinated fashion. ■

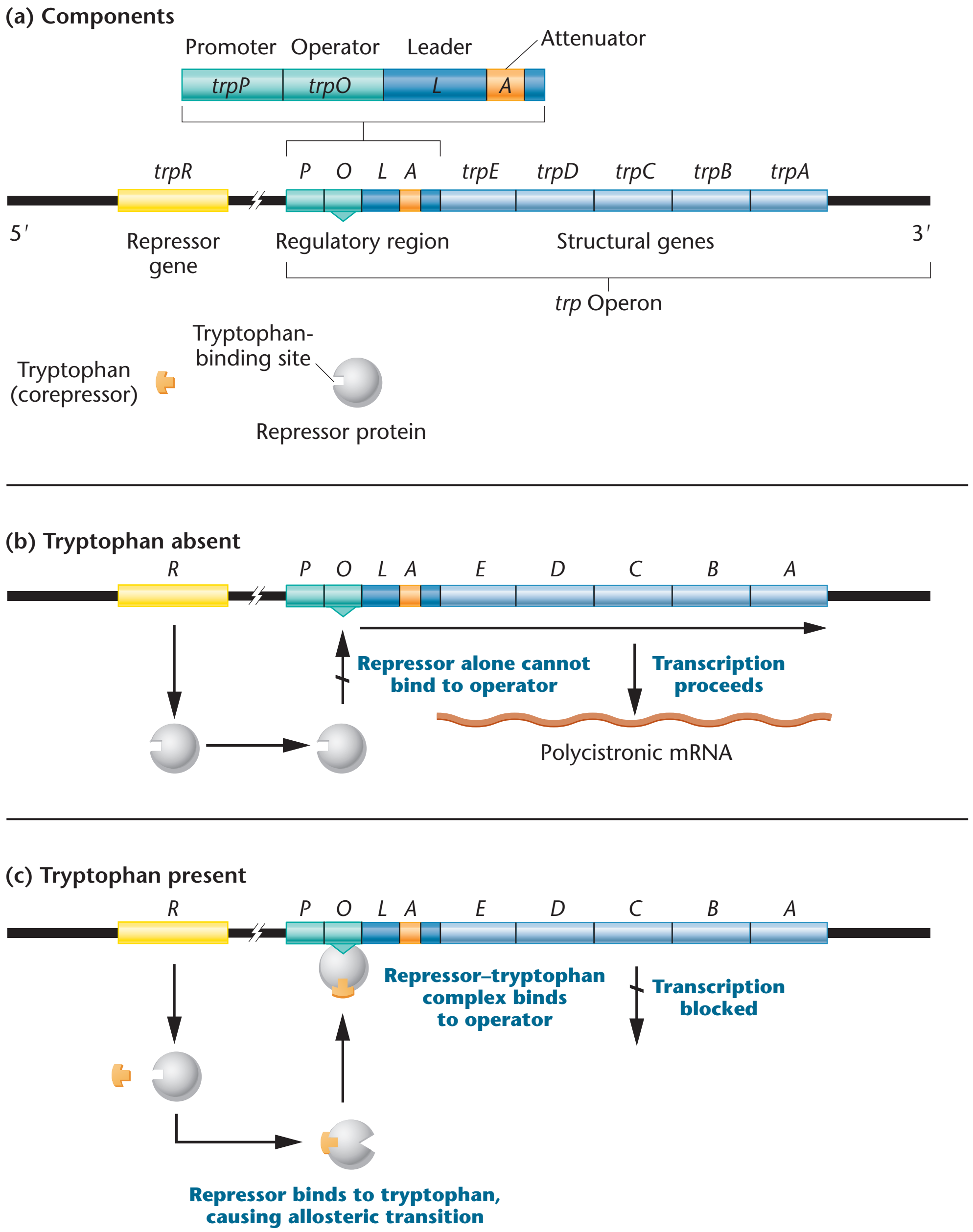


FIGURE 16.12 A repressible operon. (a) The components involved in regulation of the tryptophan operon. (b) In the absence of tryptophan, an inactive repressor is made that cannot bind to the operator (O), thus allowing transcription to proceed. (c) When tryptophan is present, it binds to the repressor, causing an allosteric transition to occur. This complex binds to the operator region, leading to repression of the operon.

16.6 RNA Plays Diverse Roles in Regulating Gene Expression in Bacteria

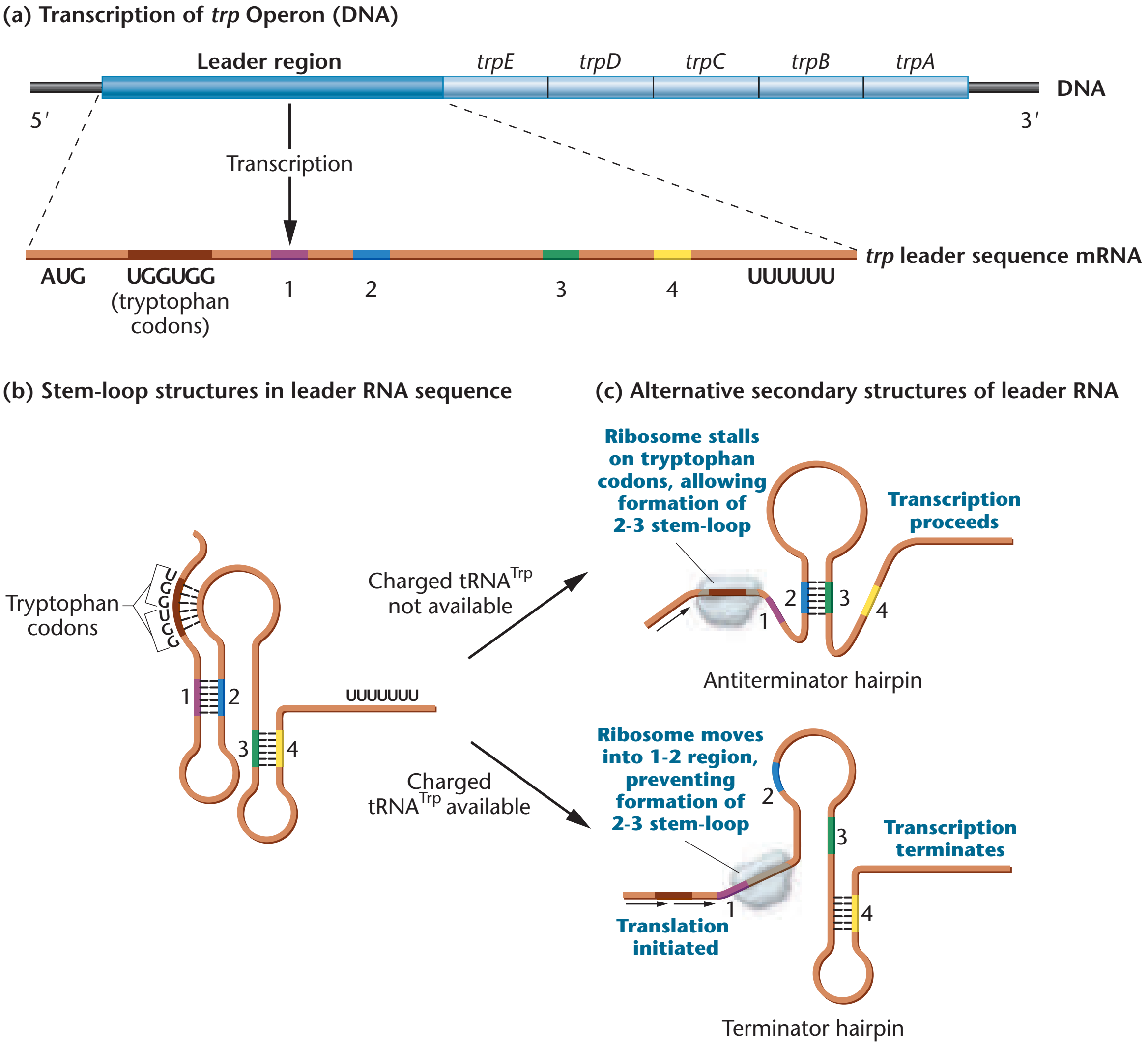
In the preceding sections of this chapter we focused on gene regulation brought about by DNA-binding regulatory proteins that interact with promoter and operator regions of the genes to be regulated. These regulatory proteins, such as the *lac* repressor and the CAP protein, act to decrease or increase transcription initiation from their target promoters by affecting the binding of RNA polymerase to the promoter.

Gene regulation in bacteria can also occur through the interactions of regulatory molecules with specific regions of a nascent mRNA, after transcription has been initiated. The binding of these regulatory molecules alters the secondary structure of the mRNA, leading to premature transcription termination or repression of translation. We will discuss three types of regulation involving RNA—*attenuation*, *riboswitches*, and *small noncoding RNAs*, abbreviated in bacteria as *sRNAs*. These types of regulation fine-tune levels of gene expression in bacteria.

Attenuation

Charles Yanofsky, Kevin Bertrand, and their colleagues observed that, when tryptophan is present and the *trp* operon is repressed, initiation of transcription still occurs at a low level but is subsequently terminated at a point about 140 nucleotides along the transcript. They called this process **attenuation,** as it "weakens or impairs" expression of the operon. In contrast, when tryptophan is absent or present in very low concentrations, transcription is initiated but is *not* subsequently terminated, instead continuing beyond the leader sequence into the structural genes.

Based on these observations, Yanofsky and colleagues presented a model to explain how attenuation occurs (**Figure 16.13**). They proposed that the initial DNA sequence that is transcribed gives rise to an mRNA sequence that has the potential to fold into two mutually exclusive stem-loop structures referred to as "hairpins." If tryptophan is scarce, an mRNA secondary structure referred to as the **antiterminator hairpin** is formed. Transcription proceeds past the antiterminator hairpin region, and the entire mRNA is subsequently produced. Alternatively, in the presence of excess tryptophan, the mRNA structure that is formed is referred to as a **terminator hairpin,** and transcription is almost always terminated prematurely, just beyond the attenuator.


(a) Transcription of *trp* Operon (DNA)

(b) Stem-loop structures in leader RNA sequence

(c) Alternative secondary structures of leader RNA

Ribosome stalls on tryptophan codons, allowing formation of 2-3 stem-loop
Transcription proceeds
1
2
3
4
Antiterminator hairpin
Ribosome moves into 1-2 region, preventing formation of 2-3 stem-loop
Transcription terminates
Translation initiated
Terminator hairpin

FIGURE 16.13 The attenuation model regulating the tryptophan operon.

A key point in Yanofsky's model is that the transcript of the leader sequence in the 5′-untranslated region (5′-UTR) of the mRNA must be translated for the antiterminator hairpin to form. This leader transcript contains two triplets (UGG) that encode tryptophan, and these are present just downstream of the initial AUG sequence that signals the initiation of translation by ribosomes. When adequate tryptophan is present, charged tRNA^{Trp} is present in the cell, whereby ribosomes translate these UGG triplets, proceed through the attenuator, and allow the *terminator hairpin* to form. The terminator hairpin signals RNA polymerase to prematurely terminate transcription, and the operon is not transcribed. If cells are starved of tryptophan, charged tRNA^{Trp} will be unavailable and ribosomes will "stall" during translation of the UGG triplets. The presence of ribosomes in this region of the mRNA interferes with the formation of the terminator hairpin but allows the formation of the antiterminator hairpin within the leader transcript. As a result, transcription proceeds, leading to expression of the entire set of structural genes.

Many other bacterial operons use attenuation to control gene expression. These include operons that encode enzymes involved in the biosynthesis of amino acids such as threonine, histidine, leucine, and phenylalanine. As with the *trp* operon, attenuation occurs in a leader sequence that contains an attenuator region.

Riboswitches

Since the elucidation of attenuation in the *trp* operon, numerous cases of gene regulation that also depend on alternative forms of mRNA secondary structure have been documented. These are examples of what are more generally referred to as **riboswitches.** As with attenuation of the *trp* operon discussed earlier, the mechanism of riboswitch regulation involves short ribonucleotide sequences (or elements) present in the 5′-UTR of mRNAs. These RNA elements are capable of binding with small molecule ligands, such as metabolites, whose synthesis or activity is controlled by the genes encoded by the mRNA. Such binding causes a conformational change in one domain of the riboswitch element, which induces another change at a second RNA domain, most often creating a transcription *terminator structure*. This terminator structure interfaces directly with the transcriptional machinery and shuts it down.

Riboswitches can recognize a broad range of ligands, including amino acids, purines, vitamin cofactors, amino sugars, and metal ions. They are widespread in bacteria. In *Bacillus subtilis,* for example, approximately 5 percent of this bacterium's genes are regulated by riboswitches. They are also found in archaea, fungi, and plants and may prove to be present in animals as well.

The two important domains within a riboswitch are the ligand-binding site, called the **aptamer,** and the **expression platform,** which is capable of forming the terminator structure. Figure 16.14 illustrates the principles involved in riboswitch control. The 5′-UTR of an mRNA is shown on the left side of the figure in the absence of the ligand (metabolite). RNA polymerase has transcribed the unbound ligand-binding site, and in the *default*

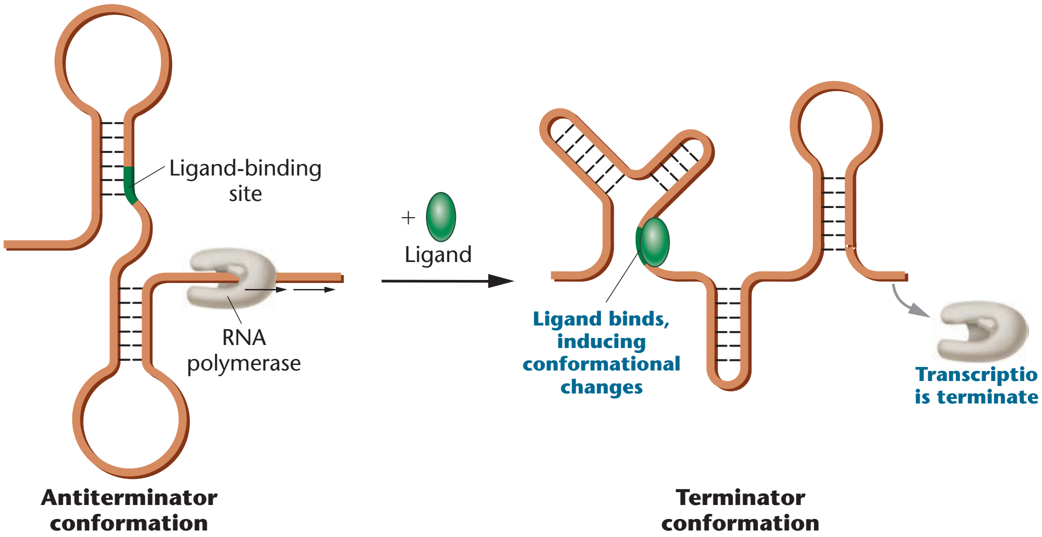

Ligand-binding site
RNA polymerase
Antiterminator conformation
+ Ligand
Ligand binds, inducing conformational changes
Transcriptio is terminate
Terminator conformation

FIGURE 16.14 An illustration of the mechanism of riboswitch regulation of gene expression, where the default position (left) is in the antiterminator conformation. Upon binding by the ligand, the mRNA adopts the terminator conformation (right).

conformation, the expression domain adopts an *antiterminator conformation*. Thus, transcription continues through the expression platform and into the coding region. On the right side of the figure, the presence of the ligand on the ligand-binding site induces an alternative conformation in the expression platform, creating the *terminator conformation*. RNA polymerase is effectively blocked and transcription ceases.

The preceding description fits the majority of riboswitches that have been studied, but there are two variations from this model that we will discuss to complete our coverage of the topic. First, it is possible for the default position of a riboswitch to be in the *terminator conformation*. For example, consider the 5′-UTRs representing genes encoding aminoacyl tRNA synthetases, the enzymes that charge tRNAs with their cognate amino acid (see Chapter 14). In these cases, uncharged tRNAs serve as the ligand and bind to the aptamer of the 5′-UTR, inducing the *antiterminator conformation*. This allows the transcription of the genes, leading to the subsequent translation of the enzymes that are required to charge tRNAs. When charged tRNAs are in abundance and uncharged tRNAs are not present to bind to the aptamer, the riboswitch, in the default condition, adapts the *terminator conformation*, effectively shutting down transcription. As predicted, charged tRNAs are unable to bind to the aptamer.

Finally, there are examples of bacterial riboswitches where, like attenuation of the *trp* operon discussed earlier, the alternative 5′-UTR conformations allow transcription to be completed, but ligand binding to the aptamer induces a conformation that inhibits translation by ribosomes. In such cases, regulation of gene expression has been invoked at the level of translation.

Small Noncoding RNAs Play Regulatory Roles in Bacteria

Bacterial **small noncoding RNAs (sRNAs)** were discovered decades ago, but their regulatory functions are still being elucidated and new sRNAs are still being discovered. It is thought that *E. coli* contain roughly 80–100 sRNAs, and other species are reported to have three times that number. sRNAs are generally between 50 and 500 nucleotides long and are involved in gene regulation and the modification of protein function. sRNAs involved in gene regulation are often transcribed from loci that partially overlap the coding genes that they regulate. However, they are transcribed from the opposite strand of DNA and in the opposite direction, making them complementary to mRNAs transcribed from that locus. In other cases, sRNAs are complementary to target mRNAs but are transcribed from loci that do not overlap target genes. sRNAs regulate gene expression by binding to mRNAs (usually at the 5′ end) that are being transcribed. In some cases, the binding of sRNAs to mRNAs blocks translation of the mRNA by masking the *ribosome-binding site* (*RBS*). In other cases, binding enhances translation by preventing secondary structures from forming in the mRNA that would block translation, often by masking the RBS (see **Figure 16.15**). Thus, sRNAs can be both negative and positive regulators of gene expression.

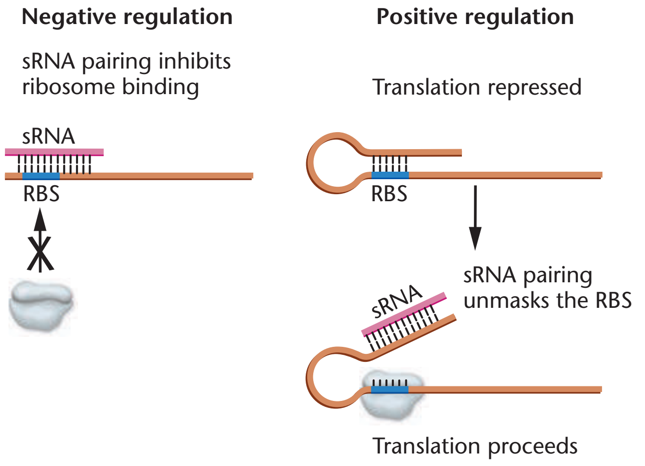


FIGURE 16.15 Bacterial small noncoding RNAs regulate gene expression. Bacterial sRNAs can be negative regulators of gene expression by binding to mRNAs and preventing translation by masking the ribosome-binding site (RBS), or they can be positive regulators of gene expression by binding to mRNAs and preventing secondary structures (that would otherwise mask an RBS) and enable translation.

sRNAs have been shown to play important roles in gene regulation in response to changing environmental conditions or stress. For example, the sRNA *DsrA* of *E. coli* is upregulated in response to low temperature and promotes the expression of genes that enable the long-term survival of the cell under stressful conditions, or in the stationary phase (see Figure 6.1). *DsrA* binds to *rpoS* mRNA to promote the translation of the RpoS stress response sigma factor, which is the primary transcriptional regulator of genes that promote the stationary phase. In contrast, *RyhB* sRNA from *E. coli* is a negative regulator of gene expression. In response to low iron levels, *RyhB* is transcribed to inhibit the translation of several nonessential iron-containing enzymes so that the more critical iron-containing enzymes can utilize what little iron is present in the cytoplasm.

CASE STUDY MRSA in the National Football League (NFL)

In 2013, there was an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) at an NFL training facility. One player suffered a career-ending infection to his foot and sued the team owners for $20 million for unsanitary conditions that contributed to the bacterial infection. A settlement with undisclosed terms was reached in 2017. MRSA is highly contagious and is spread by direct skin contact or by airborne transmission and can result in amputation or death. In addition, MRSA is very difficult to treat because it is resistant to many antibiotics. For example, β-lactam antibiotics, such as penicillin, function by binding to and inactivating bacterial penicillin-binding proteins (PBPs), which synthesize the bacterial cell wall. However, MRSA expresses an alternative type of PBP, called PBP2a encoded by the *mecA* gene. β-lactam antibiotics only weakly bind PBP2a, and thus cell wall synthesis can continue in their presence. Moreover, in a system somewhat analogous to the regulation of the *lac* operon, *mecA* is induced by the presence of β-lactam antibiotics and repressed in their absence. This "on-demand" expression of *mecA* means that when the infection is treated with antibiotics, the cells ramp up their resistance.

1. Speculate on how *mecA* expression is inducible and repressible based on what you know about the *lac* operon.
2. Chen et al. [*Antimicrob. Agents Chemother.* (2014) 58(2):1047–1054] studied several strains of MRSA isolated in Taiwan and found that some single point mutations in the *mecA* promoter were linked to increased antibiotic resistance while other point mutations were linked to decreased antibiotic resistance. Why might *mecA* promoter mutations have these opposing effects?
3. What ethical responsibility do team owners have with respect to preventing the spread of pathogenic bacteria? What responsibilities must players assume to prevent infecting other athletes?

See CDC page: Methicillin-resistant *Staphyloccocus aureus* (MRSA) (https://www.cdc.gov/mrsa/community/team-hc-providers/advice-for-athletes.html).

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Research on the *lac* operon in *E. coli* pioneered our understanding of gene regulation in bacteria.
2. Genes involved in the metabolism of lactose are coordinately regulated by a negative control system that responds to the presence or absence of lactose.
3. The catabolite-activating protein (CAP) exerts positive control over *lac* gene expression by interacting with RNA polymerase at the *lac* promoter and by responding to the levels of cyclic AMP in the bacterial cell.
4. The *lac* repressor of *E. coli* has been isolated and studied. Crystal structure analysis has shown how it interacts with the DNA of the operon as well as with inducers, revealing conformational changes in DNA leading to the formation of a repression loop that inhibits binding between RNA polymerase and the promoter region of the operon.
5. Unlike the inducible *lac* operon, the *trp* operon is repressible. In the presence of tryptophan, the repressor binds to the regulatory region of the *trp* operon and represses transcription initiation.
6. RNAs are sometimes involved in the regulation of gene expression in bacteria, including the process of attenuation, the involvement of riboswitches, and interactions involving short noncoding RNAs (sRNAs).

GENETICS, ETHICS, AND SOCIETY

Quorum Sensing: Social Networking and Gene Regulation in Bacteria

Like other social organisms, bacteria live and prosper in groups, influencing each other's behaviors, modifying their environments, and combating potential enemies. But to do this effectively, they must communicate with each other and alter their collective gene expression patterns.

One of the most successful ways that bacteria communicate is through a system known as **quorum sensing (QS)**. Bacteria that use QS are able to synthesize, secrete, and detect a number of small diffusible molecules called autoinducers (AIs). As the population of bacteria increases, the concentration of AIs also increases in their environments, such as soils, waters, and animal intestines. Once AIs reach a critical concentration, bacteria within the population begin to respond to the AI signals by regulating the expression of genes involved with functions as diverse as toxin production, biofilm formation, metabolism, and growth. In this way, bacteria can sense population size and act in ways that enhance the survival and function of the group. Quorum sensing has been described in more than 70 species of bacteria, and dozens of genes can be regulated in a bacterium in response to QS signals.

Several different types of AI molecules act as QS signals. Examples are acyl-homoserine lactones (AHLs) and

4,5-dihydroxy-2,3-pentanediones (AI-2s). Bacteria synthesize both species-specific AI molecules and AI molecules that can be detected by other species. Specific receptors that recognize these molecules can reside either in the cytoplasm of the bacteria or within membranes.

Although QS regulatory mechanisms operate within many different bacterial species, the role of QS in pathogenic bacteria is of particular interest. Many important human bacterial pathogens, such as *Staphylococcus aureus*, *Pseudomonas aeruginosa*, and *Vibrio cholerae*, use QS to regulate gene expression to enhance host invasion and avoid immune system detection during infection. For example, *Vibrio cholerae*, the causative agent of cholera, uses AI-2 and another species-specific AI to activate the genes controlling the production of cholera toxin. In this way, the pathogen can delay toxin production until the bacteria have established a high enough population density to yield effective levels of the toxin. *Pseudomonas aeruginosa*, the gram-negative bacterium that often affects patients with cystic fibrosis, uses QS to regulate the production of elastase, a protease that disrupts the respiratory epithelium and interferes with ciliary function. It also uses AIs to control the production of biofilms, tough protective shells that resist host defenses and antibiotics.

Given the widespread use of QS by pathogenic bacteria to regulate virulence in humans, QS regulatory molecules have become targets in the development of drugs to treat infections.

YOUR TURN

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand the potential applications of QS molecules in human diseases and therapies.

1. Inhibitors of quorum-sensing AI molecules have potential as antibacterial agents. What are some ways in which AI inhibitors could work to combat bacterial infections? Have any AI inhibitor therapeutics reached clinical trials? What limitations are there to using AI inhibitors?

   A discussion of AI inhibitor molecules can be found at Kalia, V. C., et al. (2014). Evolution of resistance to quorum sensing inhibitors. *Microb. Ecol*. 68:13–23.

2. Studies are revealing the importance of gut microbiota for the maintenance of health and resistance to diseases. In addition, the loss of balance between gastrointestinal microbial species is implicated in the development of chronic inflammatory bowel disease, obesity, and diabetes, as well as stress disorders and autism. Discuss how QS contributes to microbial balance in the gut and how manipulation of AI concentrations may lead to therapeutic approaches.

   Information on this topic can be found at Thompson, J. A., et al. (2015). Manipulation of the quorum sensing signal AI-2 affects the antibiotic-treated gut microbiota. *Cell Reports* 10:861–1871 *and* Azvolinsky, A. (2015). Quorum-sensing molecule modifies gut microbiota (http://www.the-scientist.com/?articles.view/articleNo/42501/title/Quorum-Sensing-Molecule-Modifies-Gut-Microbiota/).

INSIGHTS AND SOLUTIONS

A hypothetical operon (*theo*) in *E. coli* contains several structural genes encoding enzymes that are involved sequentially in the biosynthesis of an amino acid. Unlike the *lac* operon, in which the repressor gene is separate from the operon, the gene encoding the regulator molecule is contained within the *theo* operon. When the end product (the amino acid) is present, it combines with the regulator molecule, and this complex binds to the operator, repressing the operon. In the absence of the amino acid, the regulatory molecule fails to bind to the operator, and transcription proceeds.

Categorize and characterize this operon, and then consider the following mutations, as well as the situation in which the wild-type gene is present along with the mutant gene in partially diploid cells (F′):

(a) Mutation in the operator region

(b) Mutation in the promoter region

(c) Mutation in the regulator gene

In each case, will the operon be active or inactive in transcription, assuming that the mutation affects the regulation of the *theo* operon? Compare each response with the equivalent situation for the *lac* operon.

Solution: The *theo* operon is repressible and under negative control. When there is no amino acid present in the medium (or the environment), the product of the regulatory gene cannot bind to the operator region, and transcription proceeds under the direction of RNA polymerase. The enzymes necessary for the synthesis of the amino acid are produced, as is the regulator molecule. If the amino acid *is* present, either initially or after sufficient synthesis has occurred, the amino acid binds to the regulator, forming a complex that interacts with the operator region, causing repression of transcription of the genes within the operon.

The *theo* operon is similar to the tryptophan system, except that the regulator gene is within the operon rather than separate from it. Therefore, in the *theo* operon, the regulator gene is itself regulated by the presence or absence of the amino acid.

(a) As in the *lac* operon, a mutation in the *theo* operator gene inhibits binding with the repressor complex, and transcription occurs constitutively. The presence of an

(*continued*)

Insights and Solutions—continued

F′ plasmid bearing the wild-type allele would have no effect, since it is not adjacent to the structural genes.

(b) A mutation in the *theo* promoter region would no doubt inhibit binding to RNA polymerase and therefore inhibit transcription. This would also happen in the *lac* operon. A wild-type allele present in an F′ plasmid would have no effect.

(c) A mutation in the *theo* regulator gene, as in the *lac* system, may inhibit either its binding to the repressor or its binding to the operator gene. In both cases, transcription will be constitutive, because the *theo* system is repressible. Both cases result in the failure of the regulator to bind to the operator, allowing transcription to proceed. In the *lac* system, failure to bind the corepressor lactose would permanently repress the system. The addition of a wild-type allele would restore repressibility, provided that this gene was transcribed constitutively.

Problems and Discussion Questions

1. **HOW DO WE KNOW?** In this chapter, we focused on the regulation of gene expression in bacteria. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) How do we know that bacteria regulate the expression of certain genes in response to the environment?
(b) What evidence established that lactose serves as the inducer of a gene whose product is related to lactose metabolism?
(c) What led researchers to conclude that a repressor molecule regulates the *lac* operon?
(d) How do we know that the *lac* repressor is a protein?
(e) How do we know that the *trp* operon is a repressible control system, in contrast to the *lac* operon, which is an inducible control system?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on p 373. These all relate to the regulation of gene expression in bacteria. Write a brief essay that discusses why you think regulatory systems evolved in bacteria (i.e., what advantages do regulatory systems provide to these organisms?), and, in the context of regulation, discuss why genes related to common functions are found together in operons.

3. Contrast positive versus negative control of gene expression.

4. Contrast the role of the repressor in an inducible system and in a repressible system.

5. For the *lac* genotypes shown in the following table, predict whether the structural genes (*Z*) are constitutive, permanently repressed, or inducible in the presence of lactose.

Genotype	Constitutive	Repressed	Inducible
$I^+O^+Z^+$			×
$I^-O^+Z^+$			
$I^-O^CZ^+$			
$I^-O^CZ^+/F'O^+$			
$I^+O^CZ^+/F'O^+$			
$I^SO^+Z^+$			
$I^SO^+Z^+/F'I^+$			

6. For the genotypes and conditions (lactose present or absent) shown in the following table, predict whether functional enzymes, nonfunctional enzymes, or no enzymes are made.

Genotype	Condition	Functional Enzyme Made	Nonfunctional Enzyme Made	No Enzyme Made
$I^+O^+Z^+$	No lactose			×
$I^+O^CZ^+$	Lactose			
$I^-O^+Z^-$	No lactose			
$I^-O^+Z^-$	Lactose			
$I^-O^+Z^+/F'I^+$	No lactose			
$I^+O^CZ^+/F'O^+$	Lactose			
$I^+O^+Z^-/F'I^+O^+Z^+$	Lactose			
$I^-O^+Z^-/F'I^+O^+Z^+$	No lactose			
$I^SO^+Z^+/F'O^+$	No lactose			
$I^+O^CZ^+/F'O^+Z^+$	Lactose			

7. The locations of numerous $lacI^-$ and $lacI^S$ mutations have been determined within the DNA sequence of the *lacI* gene. Among these, $lacI^-$ mutations were found to occur in the 5′-upstream region of the gene, while $lacI^S$ mutations were found to occur farther downstream in the gene. Are the locations of the two types of mutations within the gene consistent with what is known about the function of the repressor that is the product of the *lacI* gene?

8. Describe the experimental rationale that allowed the *lac* repressor to be isolated.

9. What properties demonstrate that the *lac* repressor is a protein? Describe the evidence that it indeed serves as a repressor within the operon system.

10. Predict the effect on the inducibility of the *lac* operon of a mutation that disrupts the function of (a) the *crp* gene, which encodes the CAP protein, and (b) the CAP-binding site within the promoter.

11. Erythritol, a natural sugar abundant in fruits and fermenting foods, is about 65 percent as sweet as table sugar and has about

95 percent fewer calories. It is "tooth friendly" and generally devoid of negative side effects as a human consumable product. Pathogenic *Brucella* strains that catabolize erythritol contain four closely spaced genes, all involved in erythritol metabolism. One of the four genes (*eryD*) encodes a product that represses the expression of the other three genes. Erythritol catabolism is stimulated by erythritol. Present a simple regulatory model to account for the regulation of erythritol catabolism in *Brucella*. Does this system appear to be under inducible or repressible control?

12. Describe the role of attenuation in the regulation of tryptophan biosynthesis.

13. Attenuation of the *trp* operon was viewed as a relatively inefficient way to achieve genetic regulation when it was first discovered in the 1970s. Since then, however, attenuation has been found to be a relatively common regulatory strategy. Assuming that attenuation is a relatively inefficient way to achieve genetic regulation, what might explain its widespread occurrence?

14. Neelaredoxin is a 15-kDa protein that is a gene product common in anaerobic bacteria. It has superoxide-scavenging activity, and it is *constitutively expressed*. In addition, its expression is not further *induced* during its exposure to O_2 or H_2O_2 [Silva, G. et al. (2001). *J. Bacteriol*. 183:4413–4420]. What do the terms *constitutively expressed* and *induced* mean in terms of neelaredoxin synthesis?

15. The creation of milk products such as cheeses and yogurts is dependent on the conversion by various anaerobic bacteria, including several *Lactobacillus* species, of lactose to glucose and galactose, ultimately producing lactic acid. These conversions are dependent on both permease and β-galactosidase as part of the *lac* operon. After selection for rapid fermentation for the production of yogurt, one *Lactobacillus* subspecies lost its ability to regulate *lac* operon expression [Lapierre, L., et al. (2002). *J. Bacteriol*. 184:928–935]. Would you consider it likely that in this subspecies the *lac* operon is on or off? What genetic events would likely contribute to the loss of regulation as described above?

16. Both attenuation of the *trp* operon in *E. coli* and riboswitches in *B. subtilis* rely on changes in the secondary structure of the leader regions of mRNA to regulate gene expression. Compare and contrast the specific mechanisms in these two types of regulation with that involving short noncoding RNAs (sRNAs).

17. Keeping in mind the life cycle of bacteriophages discussed earlier in the text (see Chapter 6), consider the following problem: During the reproductive cycle of a temperate bacteriophage, the viral DNA inserts into the bacterial chromosome where the resultant prophage behaves much like a Trojan horse. It can remain quiescent, or it can become lytic and initiate a burst of progeny viruses. Several operons maintain the prophage state by interacting with a repressor that keeps the lytic cycle in check. Insults (ultraviolet light, for example) to the bacterial cell lead to a partial breakdown of the repressor, which in turn causes the production of enzymes involved in the lytic cycle. As stated in this simple form, would you consider this system of regulation to be operating under positive or negative control?

Extra-Spicy Problems

18. Bacterial strategies to evade natural or human-imposed antibiotics are varied and include membrane-bound efflux pumps that export antibiotics from the cell. A review of efflux pumps [Grkovic, S., et al. (2002)] states that, because energy is required to drive the pumps, activating them in the absence of the antibiotic has a selective disadvantage. The review also states that a given antibiotic may play a role in the regulation of efflux by interacting with either an activator protein or a repressor protein, depending on the system involved. How might such systems be categorized in terms of *negative control* (*inducible* or *repressible*) or *positive control* (*inducible* or *repressible*)?

19. In a theoretical operon, genes *A*, *B*, *C*, and *D* represent the repressor gene, the promoter sequence, the operator gene, and the structural gene, *but not necessarily in the order named*. This operon is concerned with the metabolism of a theoretical molecule (tm). From the data provided in the accompanying table, first decide whether the operon is inducible or repressible. Then assign *A*, *B*, *C*, and *D* to the four parts of the operon. Explain your rationale. (AE = active enzyme; IE = inactive enzyme; NE = no enzyme.)

Genotype	tm Present	tm Absent
$A^+B^+C^+D^+$	AE	NE
$A^-B^+C^+D^+$	AE	AE
$A^+B^-C^+D^+$	NE	NE
$A^+B^+C^-D^+$	IE	NE
$A^+B^+C^+D^-$	AE	AE
$A^-B^+C^+D^+/F'A^+B^+C^+D^+$	AE	AE
$A^+B^-C^+D^+/F'A^+B^+C^+D^+$	AE	NE
$A^+B^+C^-D^+/F'A^+B^+\ C^+D^+$	AE + IE	NE
$A^+B^+C^+D^-/F'A^+B^+C^+D^+$	AE	NE

20. A bacterial operon is responsible for the production of the biosynthetic enzymes needed to make the hypothetical amino acid tisophane (tis). The operon is regulated by a separate gene, *R*. The deletion of *R* causes the loss of enzyme synthesis. In the wild-type condition, when tis is present, no enzymes are made; in the absence of tis, the enzymes are made. Mutations in the operator gene (O^-) result in repression regardless of the presence of tis. Is the operon under positive or negative control? Propose a model for (a) repression of the genes in the presence of tis in wild-type cells and (b) the mutations.

21. A marine bacterium is isolated and shown to contain an inducible operon whose genetic products metabolize oil when it is encountered in the environment. Investigation demonstrates that the operon is under positive control and that there is a *reg* gene whose product interacts with an operator region (*o*) to regulate the structural genes, designated *sg*.

In an attempt to understand how the operon functions, a constitutive mutant strain and several partial diploid strains were isolated and tested with the results shown in the following table.

Host Chromosome	F′ Factor	Phenotype
Wild type	None	Inducible
Wild type	*reg* gene from mutant strain	Inducible
Wild type	Operon from mutant strain	Constitutive
Mutant strain	*reg* gene from wild type	Constitutive

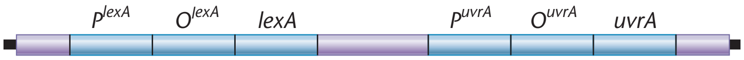
Draw all possible conclusions about the mutation as well as the nature of regulation of the operon. Is the constitutive mutation in the *trans*-acting *reg* element or in the *cis*-acting *o* operator element?

22. The SOS repair genes in *E. coli* (discussed in Chapter 15) are negatively regulated by the *lexA* gene product, called the LexA repressor. When a cell's DNA sustains extensive damage, the LexA repressor is inactivated by the *recA* gene product (RecA), and transcription of the SOS genes is increased dramatically.

One of the SOS genes is the *uvrA* gene. You are a student studying the function of the *uvrA* gene product in DNA repair. You isolate a mutant strain that shows constitutive expression of the UvrA protein. Naming this mutant strain *uvrA*^{C}, you construct the diagram shown above in the right-hand column showing the *lexA* and *uvrA* operons:

(a) Describe two different mutations that would result in a *uvrA* constitutive phenotype. Indicate the actual genotypes involved.

(b) Outline a series of genetic experiments that would use partial diploid strains to determine which of the two possible mutations you have isolated.



23. A fellow student considers the issues in Problem 22 and argues that there is a more straightforward, nongenetic experiment that could differentiate between the two types of mutations. The experiment requires no fancy genetics and would allow you to easily assay the products of the other SOS genes. Propose such an experiment.

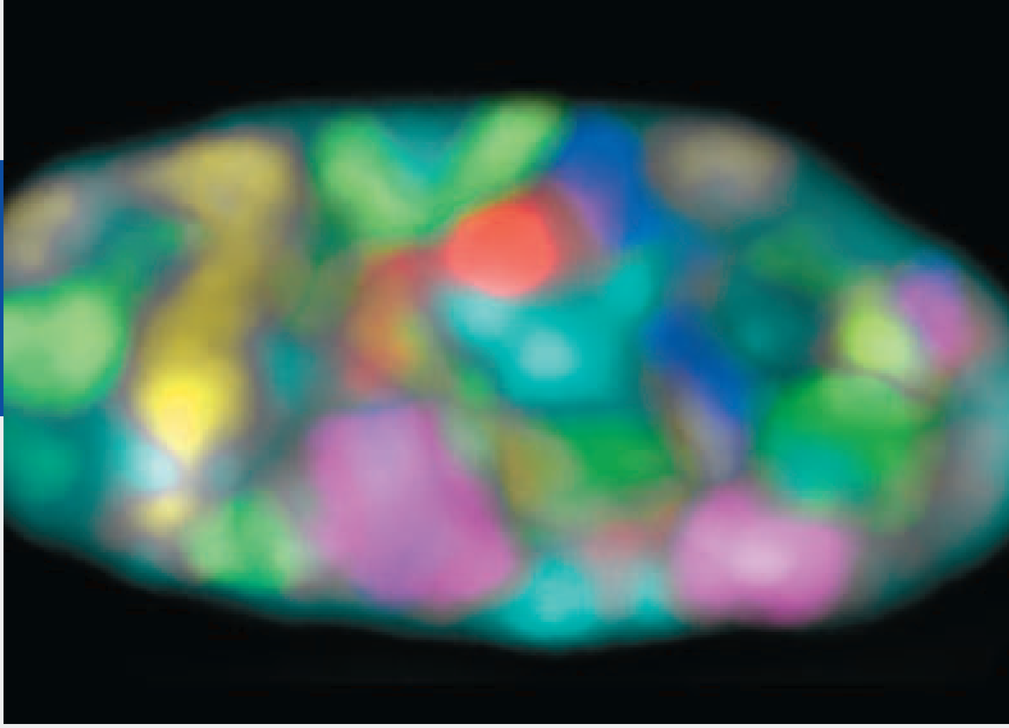
24. Figure 16.13 depicts numerous critical regions of the leader sequence of mRNA that play important roles during the process of attenuation in the *trp* operon. A closer view of the leader sequence, which begins at about position 30 downstream from the 5′ end, is shown below, running along both columns.

Within this molecule are the sequences that cause the formation of the alternative hairpins. It also contains the successive triplets that encode tryptophan, where stalling during translation occurs. Take a large piece of paper (such as manila wrapping paper) and, along with several other students from your genetics class, work through the base sequence to identify the *trp* codons and the parts of the molecule representing the base-pairing regions that form the terminator and antiterminator hairpins shown in Figure 16.13.

AUGAAAGCAAUUUUCGUACUGAAAGGUUGGUGGCGCACUUCCUGAAACGGGCAGUGUAUUCACCAUGCGUAAAGCAAUCAGAUACCCAGCCCGCCUAAUGAGCGGGCUUUUUUUU

(Leader sequence for Problem 24 above)

17



Transcriptional Regulation in Eukaryotes

Chromosome territories in a human fibroblast cell nucleus. Each chromosome is stained with a different-colored probe.

CHAPTER CONCEPTS

- While transcription and translation are tightly coupled in bacteria, in eukaryotes, these processes are spatially and temporally separated, and thus independently regulated.
- Gene expression is regulated at many steps in eukaryotes, including DNA packaging into chromatin, transcription, RNA processing, RNA export from the nucleus, RNA stability, RNA localization, translation, and posttranslational modifications.
- Chromatin remodeling, as well as modifications to DNA and histones, play important roles in regulating gene expression in eukaryotes.
- Eukaryotic transcription initiation requires the assembly of transcription regulatory proteins at *cis*-acting DNA sites known as promoters, enhancers, and silencers.
- Transcriptional activators and repressors influence the association of the general transcription factors into pre-initiation complexes at gene promoters, and they may influence chromatin remodeling or modifications.
- The genome-wide identification of transcription factor binding sites and chromatin modifications in the human genome in a manner that is cell-type specific has proven to be a powerful tool for understanding transcriptional regulation and how it relates to disease.

Virtually all cells in a multicellular eukaryotic organism contain a complete genome; however, such organisms often possess different cell types with diverse morphologies and physiological functions. This simple observation highlights the importance of the regulation of gene expression in eukaryotes. For example, skin cells and muscle cells differ in appearance and function because they express different genes. Skin cells express keratins, fibrous structural proteins that bestow the skin with protective properties. Muscle cells express high levels of myosin II, a protein that mediates muscle contraction. Skin cells do not express myosin II, and muscle cells do not express keratins.

In addition to gene expression that is cell-type specific, some genes are only expressed under certain conditions or at certain times. For example, when oxygen levels in the blood are low, such as at high altitude or after rigorous exercise, expression of the hormone erythropoietin is upregulated, which leads to an increase in red blood cell production and thus oxygen-carrying capacity.

Underscoring the importance of regulation, the misregulation of genes in eukaryotes is associated with developmental defects and disease. For instance, the overexpression of genes that regulate cellular growth can lead to uncontrolled cellular proliferation, a hallmark of cancer.

Therefore, understanding the mechanisms that control gene expression in eukaryotes is of great interest and may lead to therapies for human diseases.

We will start this chapter by briefly comparing and contrasting eukaryotic gene expression with that of bacteria to highlight several additional levels of regulation of gene expression in eukaryotes. We will then shift our focus to the regulation of transcription in eukaryotes—the first and most important step in gene expression. While also very important in eukaryotes, posttranscriptional gene regulatory mechanisms are sufficiently complex that they will be discussed in more detail later in the text (see Chapter 18).

17.1 Organization of the Eukaryotic Cell Facilitates Gene Regulation at Several Levels

As you've previously seen (Chapter 16), in bacteria, the regulation of gene expression is often linked to the metabolic needs of the cell. For example, bacteria express genes to metabolize lactose when it is present in the environment. Gene expression in bacteria is largely controlled by mechanisms that exert positive or negative control over transcription and translation (see Chapter 16). While positive and negative regulation of transcription and translation are prominent regulatory mechanisms in eukaryotes as well, there are many important differences between these processes to consider, and several additional levels for gene regulation are found in eukaryotes. **Figure 17.1** compares differences in the main mechanisms of gene regulation in bacteria and eukaryotes, which include the following:

- Eukaryotic DNA, unlike that of bacteria, is associated with histones and other proteins to form chromatin. Eukaryotic cells decrease chromatin compaction to make genes accessible to transcription and increase chromatin compaction to inhibit transcription.
- In bacteria, the processes of transcription and translation both take place in the cytoplasm and are coupled. However, in eukaryotes the organization of DNA in the nucleus and ribosomes in the cytoplasm means that transcription and translation are separated spatially and temporally.
- Whereas bacterial mRNAs are translated directly, the mRNAs of many eukaryotic genes must be spliced, capped, and polyadenylated prior to export from the

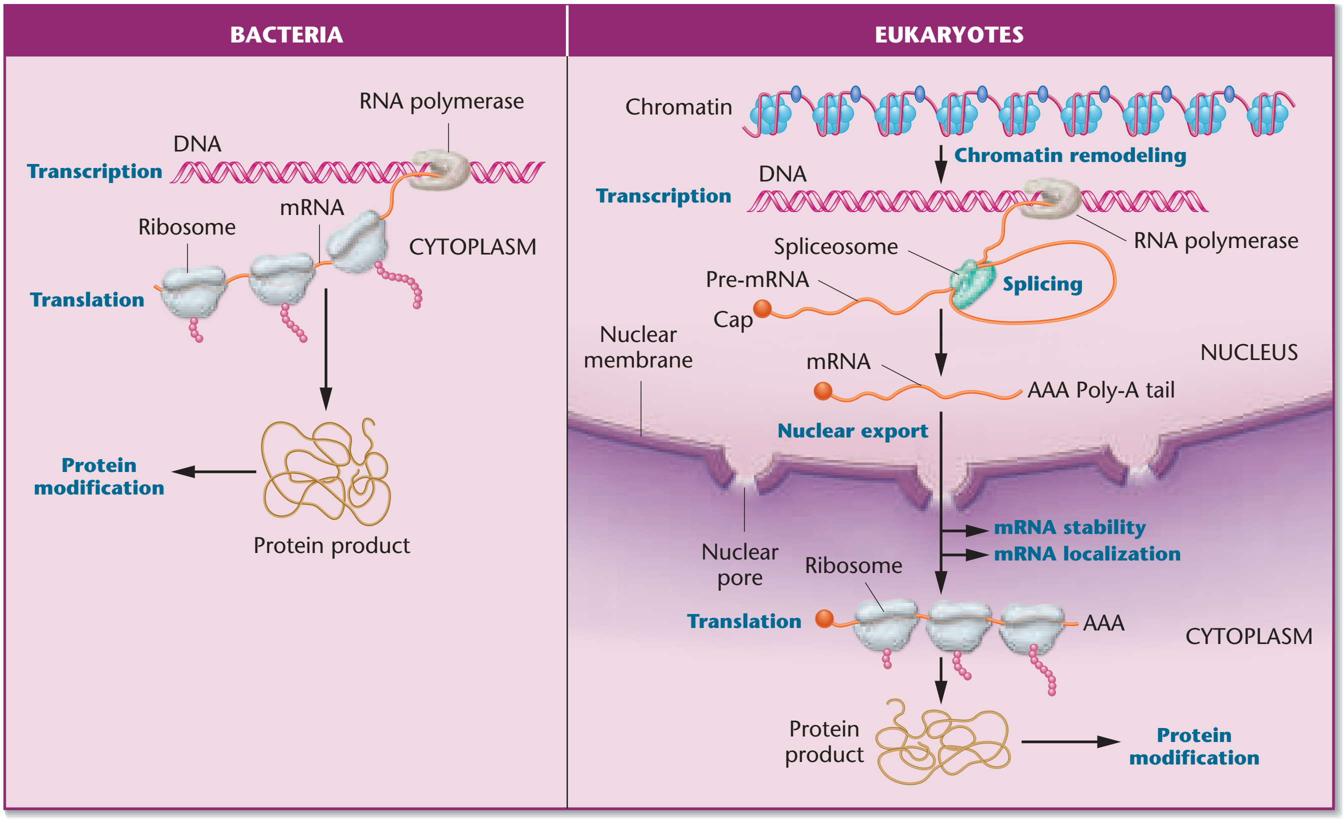


FIGURE 17.1 A comparison of gene regulation in bacteria (left) and eukaryotes (right).

nucleus and translation in the cytoplasm. Each of these processes can be regulated in order to influence the numbers and types of mRNAs available for translation.

- Bacterial mRNAs are often translated immediately and degraded very rapidly. In contrast, eukaryotic mRNAs can have long half-lives and can be transported, localized, and translated in specific subcellular destinations.
- Proteins in bacteria and in eukaryotes can be posttranslationally modified by processes such as phosphorylation and methylation, which serve many functions, including the regulation of protein activities. However, the repertoire of posttranslational modifications in eukaryotes is more extensive, leading to additional regulatory opportunities.

The focus of this chapter is mechanisms of transcriptional regulation. (Posttranscriptional mechanisms will be covered in Chapter 18.) As discussed earlier in the text (see Chapter 13), eukaryotes use three different RNA polymerases (RNAPs) for transcription. RNAP II transcribes protein-coding genes and some noncoding RNAs. In contrast, RNA polymerases I and III transcribe genes that code for ribosomal RNAs, transfer RNAs, and some small nuclear RNAs. The genes transcribed by RNA polymerase I and III are regulated differently than those transcribed by RNAP II. The promoters recognized by each type of RNA polymerase have different nucleotide sequences and bind different transcription factors. In addition, genes transcribed by each polymerase have different transcription termination signals and RNA processing mechanisms. For simplicity, we will limit our discussion to regulation of genes transcribed by RNAP II, presently the best studied of the eukaryotic RNAPs.

17.2 Eukaryotic Gene Expression Is Influenced by Chromatin Modifications

Recall from earlier in the text (see Chapter 12) that eukaryotic DNA is combined with histones and nonhistone proteins to form **chromatin.** The basic structure of chromatin is characterized by repeating DNA–histone complexes called **nucleosomes** that are wound into 30-nm fibers, which in turn form other, even more compact structures. The presence of these compact chromatin structures inhibits many processes, including DNA replication, repair, and transcription. In this section, we outline some of the ways in which eukaryotic cells modify chromatin in order to regulate gene expression.

Chromosome Territories and Transcription Factories

Despite the widespread analogy that chromosomes within the nucleus look like a bowl of cooked spaghetti, the nucleus has an elegant architecture. Chromosome-labeling techniques have revealed that each chromosome within the interphase nucleus occupies a discrete domain called a **chromosome territory** and stays separate from other chromosomes (see the chapter-opening image on page 393). Channels between chromosomes contain little or no DNA and are called **interchromatin compartments.**

Research suggests that transcriptionally active genes are located at the edges of chromosome territories next to the channels of the interchromatin compartments. This organization brings actively transcribed genes into closer association with each other and with the transcriptional machinery, thereby facilitating their coordinated expression. Transcripts produced at the edge of chromosome territories move into the adjacent interchromatin compartments, which house RNA processing machinery and are contiguous with nuclear pores. This arrangement facilitates the capping, splicing, and poly-A tailing of mRNAs during and after transcription, and the eventual export of mRNAs into the cytoplasm.

Keeping in mind this view of nuclear architecture, if we zoom in at the edge of a chromosome territory, we find another important organizational feature within the nucleus—the **transcription factory** (Figure 17.2). Transcription factories are nuclear sites that are estimated to contain ~4–30 active RNA polymerase molecules and transcription regulatory molecules. There is mounting evidence that some transcription

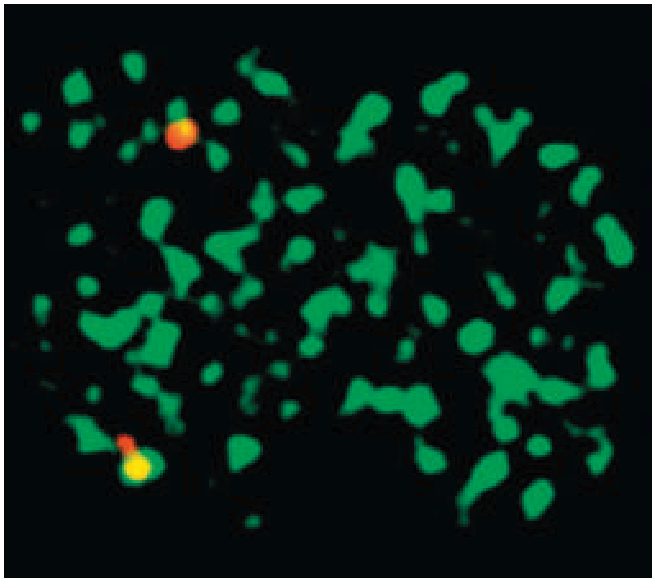


FIGURE 17.2 Visualization of active sites of transcription, known as transcription factories (green), in mouse cell nuclei. Transcription of one specific gene (at both alleles), encoding the cell-cycle regulatory protein Myc, is visualized with a sequence-specific probe (red; red and green colocalization is yellow).

factories transcribe genes that are regulated by the same transcriptional activators. In this way, transcription factories impact nuclear organization by clustering coregulated genes. Supporting this view, live imaging shows that transcription factories are dynamic structures that can form rapidly, then disassemble, as transcription is stimulated and repressed. In addition, evidence suggests that the number of transcription factories in a nucleus varies from ~100–8000 depending on the specific cell type, reflecting the different transcriptional demands of different cells.

It is clear that nuclear architecture and transcriptional regulation are interdependent; changes in nuclear architecture affect transcription, and changes in transcriptional activity necessitate changes in chromosome organization. However, there are still many unanswered questions about how these two processes are coordinated. To address this, the U.S. National Institutes of Health launched a new project in 2014 called the 4D Nucleome with the stated goal of determining "the role nuclear organization plays in gene expression and cellular function, and how changes in the nuclear organization affect normal development as well as various diseases."

Open and Closed Chromatin

The tight association of DNA with histones and other chromatin-binding proteins inhibits access to the DNA by proteins involved in many functions including transcription. This inhibitory conformation is often referred to as "closed" chromatin. Before transcription can be initiated on nucleosomal DNA, the structure of chromatin must become "open" to RNA polymerase.

One of the first demonstrations of a link between open chromatin and active transcription was reported in 1976 in a study by Harold Weintraub and Mark Groudine. In this study, Weintraub and Groudine isolated nuclei from chicken red blood cells, fibroblasts, and brain cells. They then treated the nuclei with low levels of the enzyme DNase I. This enzyme digests DNA to varying degrees, depending on the chromatin's conformation. DNA within fully compacted chromatin is shielded or protected from DNase I, while DNA located in regions of more relaxed chromatin is rapidly digested.

After treating the nuclei with DNase I, they isolated DNA from the nuclei and probed it for the presence of sequences from three genes: β-globin (transcribed only in red blood cells), avian tumor virus (transcribed in all three cell types), and ovalbumin (not transcribed in any of these cell types). They found that β-globin DNA was digested in red blood cell nuclei treated with DNase I, but not in treated nuclei of fibroblasts or brain cells. In contrast, the avian tumor virus DNA was digested in nuclei treated with DNase I from all three cell types, and ovalbumin DNA was not digested in any of the cell nuclei. From these studies, Weintraub and Groudine concluded that while both transcriptionally active and transcriptionally inactive genes are associated with nucleosomes, the nucleosomes associated with transcriptionally active genes are in an altered conformation that renders their DNA sensitive to DNase I digestion.

Since Weintraub and Groudine's study, scientists have defined more clearly the nature of the altered chromatin conformations that are associated with transcriptionally active and inactive states. These include histone modifications, chromatin remodeling, and DNA modifications such as methylation.

Histone Modifications and Chromatin Remodeling

One way in which nucleosomes can be modified is by changing their histone composition. For example, most nucleosomes contain histone H2A (see Table 12.2). However, nucleosomes containing variant histones, such as H2A.Z, influence transcription. Whereas nucleosomes are generally a physical barrier to RNA polymerases and DNA-binding transcriptional regulators, nucleosomes containing the H2A.Z variant are not as stable and thus are less of a barrier. Interestingly, H2A.Z-containing nucleosomes are enriched at gene regulatory sequences, such as the promoters and enhancers, of actively expressed genes. This suggests that the histone composition of nucleosomes can influence transcription.

A second mechanism of chromatin alteration involves histone modification. Histone modification refers to the covalent addition of functional groups to the N-terminal tails of histone proteins. The most common histone modifications are added acetyl, methyl, or phosphate groups.

Acetylation decreases the positive charge on histones, resulting in a reduced affinity of the histone for the negative charges on the backbone phosphates of DNA. This in turn may assist the formation of open chromatin conformations, which would allow the binding of transcription regulatory proteins to DNA.

Histone acetylation is catalyzed by **histone acetyltransferase (HAT)** enzymes. In some cases, HATs are recruited to genes by the presence of certain transcriptional activator proteins that bind to transcription regulatory regions. In other cases, transcriptional activator proteins themselves have HAT activity. Of course, what can be opened can also be closed. In that case, **histone deacetylases (HDACs)** remove acetyl groups from histone tails. HDACs can be recruited to genes by some transcriptional repressor proteins that bind to gene regulatory sequences.

A third alteration mechanism is chromatin remodeling, which involves the repositioning or removal of nucleosomes on DNA, brought about by **chromatin remodeling complexes.** Chromatin remodeling complexes are large multi-subunit enzymes that use the energy of ATP hydrolysis to move and rearrange nucleosomes. Repositioning of nucleosomes makes regions of the chromosome accessible to transcription regulatory proteins, such as transcriptional activators, and RNAP II.

One of the best-studied remodeling complexes is **SWI/SNF.** Remodelers such as SWI/SNF can act in several different ways (Figure 17.3). They may loosen the association between histones and DNA, resulting in the nucleosome sliding along the DNA and exposing regulatory regions. Alternatively, they may loosen the DNA strand from the nucleosome core, or they may cause reorganization of the internal nucleosome components such as swapping in and out histone variants. In all cases, the DNA is left transiently exposed to transcription factors and RNA polymerase. Like HATs, chromatin remodeling complexes can be recruited to specific genes by transcriptional activator proteins that bind to regulatory sequences.

DNA Methylation

Another type of chromatin modification that plays a role in gene regulation is the enzyme-mediated addition or removal of methyl groups to or from bases in DNA. **DNA methylation** in eukaryotes most often occurs at position 5 of cytosine **(5-methylcytosine),** causing the methyl group to protrude into the major groove of the DNA helix. Methylation occurs most often on the cytosine of CG doublets in DNA, usually on both strands:

$$5'\text{–}{}^{m}\text{CpG–}3'$$

$$3'\text{–GpC}^{m}\text{–}5'$$

Methylatable CpG sequences are not randomly distributed throughout the genome, but tend to be concentrated in CpG-rich regions, called **CpG islands,** which are often located in or near promoter regions. Roughly 70 percent of human genes have a CpG island in their promoter sequence.

Evidence of a role for methylation in eukaryotic gene expression is based on several observations. First, an inverse relationship exists between the degree of methylation and the degree of expression. Large transcriptionally inert regions of the genome, such as the inactivated X chromosome in female mammalian cells (see Chapter 7), are often heavily methylated.

Second, methylation patterns are tissue specific and, once established, are maintained for all cells of that tissue. It appears that proper patterns of DNA methylation are essential for normal mammalian development. Despite this somatic maintenance, DNA methylation in specific regions can be altered by methylase and demethylase enzymes in order to silence or activate the transcription of genes in such regions.

Perhaps the most direct evidence of a role for methylation in gene expression comes from studies using base analogs. For example, the nucleoside **5-azacytidine** can be incorporated into DNA in place of cytidine during DNA replication. This analog cannot be methylated, causing the undermethylation of the sites where it is incorporated. The incorporation of 5-azacytidine into DNA changes the pattern of gene expression and stimulates expression of alleles on inactivated X chromosomes. In addition, the presence of 5-azacytidine in DNA can induce the expression of genes that would normally be silent in certain differentiated cells.

By what mechanism might methylation affect gene regulation? Data from *in vitro* studies suggest that methylation can repress transcription by inhibiting the binding of transcription factors to DNA. Methylated DNA may also recruit repressive chromatin remodeling complexes and HDACs to gene-regulatory regions.

It is important to know that while cytosine methylation is clearly an important mechanism for regulating gene expression in some eukaryotes, it is not uniformly true for all eukaryotes. In fact, the extent of cytosine methylation in the genome varies across species: 14 percent of cytosines are methylated in the genome of *Arabidopsis thaliana* (a flowering plant), 4 percent in that of mice, and 1 percent in that of humans, while DNA methylation is absent in the roundworm *Caenorhabditis elegans*. This suggests that while some

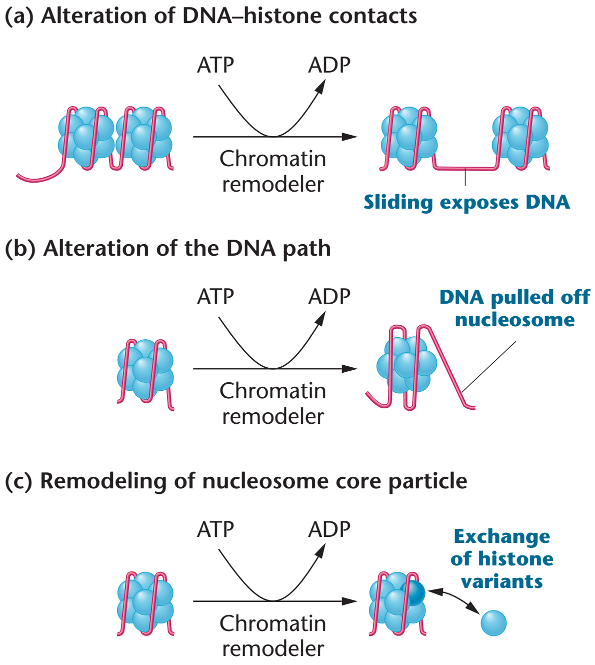

(a) Alteration of DNA–histone contacts
ATP
ADP
Chromatin remodeler
Sliding exposes DNA
(b) Alteration of the DNA path
ATP
ADP
Chromatin remodeler
DNA pulled off nucleosome
(c) Remodeling of nucleosome core particle
ATP
ADP
Chromatin remodeler
Exchange of histone variants

FIGURE 17.3 Three ways by which chromatin remodelers, such as the SWI/SNF complex, alter the association of nucleosomes with DNA. (a) The DNA–histone contacts may be loosened, allowing the nucleosomes to slide along the DNA, exposing DNA regulatory regions. (b) The path of the DNA around a nucleosome core particle may be altered. (c) Components of the core nucleosome particle may be rearranged, such as by swapping in and out variant histone proteins.

eukaryotic species have extensive methylation-based gene regulation, others do not employ this mechanism at all.

NOW SOLVE THIS

17.1 Cancer cells often have abnormal patterns of chromatin modifications. In some cancers, the DNA repair genes *MLH1* and *BRCA1* are hypermethylated on their promoter regions. Explain how this abnormal methylation pattern could contribute to cancer.

- **HINT:** *This problem involves an understanding of the types of genes that are mutated in cancer cells. The key to its solution is to consider how DNA methylation affects gene expression of cancer-related genes.*

17.3 Eukaryotic Transcription Initiation Requires Specific *Cis*-Acting Sites

Regulation of eukaryotic transcription requires the binding of many regulatory factors to specific DNA sequences located in and around genes, as well as to sequences located at great distances. Although these sequences do not, by themselves, regulate transcription, they are essential because they position regulatory proteins in regions where those proteins can act to stimulate or repress transcription of the associated gene. In this section, we will discuss some of the DNA sequences—known as *cis*-acting elements—that are required for the accurate and regulated transcription of genes transcribed by RNAP II. As defined earlier in the text (see Chapter 13), ***cis*-acting DNA elements** are located on the same chromosome as the gene that they regulate. This is in contrast to ***trans*-acting factors** (such as DNA-binding proteins or small RNA molecules) that can influence the expression of a gene on any chromosome.

Promoters and Promoter Elements

A **promoter** is a region of DNA that is recognized and bound by the basic transcriptional machinery—including RNAP II and the general transcription factors (see Chapter 13). Promoters are required for transcription initiation and are located immediately adjacent to the genes they regulate. They specify the site or sites at which transcription begins (the **transcription start site**), as well as the direction of transcription along the DNA.

There are two subcategories within eukaryotic promoters. First, **core promoters** are the minimum part of the promoter needed for accurate initiation of transcription by RNAP II. Core promoters are sequences that are ~80 nucleotides long and include the transcription start site. Second, **proximal promoter elements** are generally located up to ~250 nucleotides upstream of the transcription start site and contain binding sites for sequence-specific DNA-binding proteins that modulate the efficiency of transcription.

Recent bioinformatic and genomic research has introduced new complexities to our understanding of promoters and how they work. This research reveals that there is a great deal of diversity in eukaryotic promoters, in terms of both their structures and functions. Core promoters are classified in two ways with respect to transcription start sites. **Focused core promoters** specify transcription initiation at a single specific start site. In contrast, **dispersed core promoters** direct initiation from a number of weak transcription start sites located over a 50- to 100-nucleotide region (**Figure 17.4**).

The major type of initiation for most genes of lower eukaryotes is focused transcription initiation, while about 70 percent of vertebrate genes employ dispersed promoters. Focused promoters are usually associated with genes whose transcription levels are highly regulated in terms of time or place. Dispersed promoters, in contrast, are associated with genes that are transcribed constitutively, so-called *housekeeping genes* whose expression is required in almost all cell types. Thus, a single transcription start site may facilitate precise regulation of some genes, whereas multiple start sites may allow for a steady level of transcription of genes that are required constitutively.

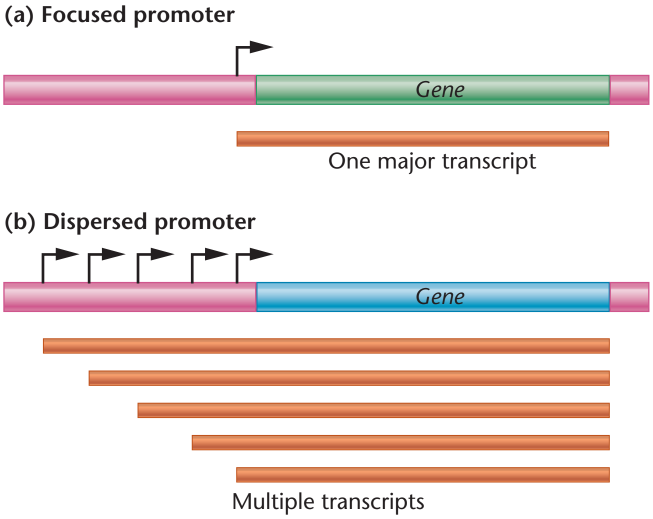


FIGURE 17.4 Focused and dispersed promoters. Focused promoters (a) specify one specific transcription initiation site. Dispersed promoters (b) specify weak transcription initiation at multiple start sites over an approximately 100-bp region. Dispersed promoters are common in vertebrates and are associated with housekeeping genes. Transcription start sites and the directions of transcription are indicated with arrows.

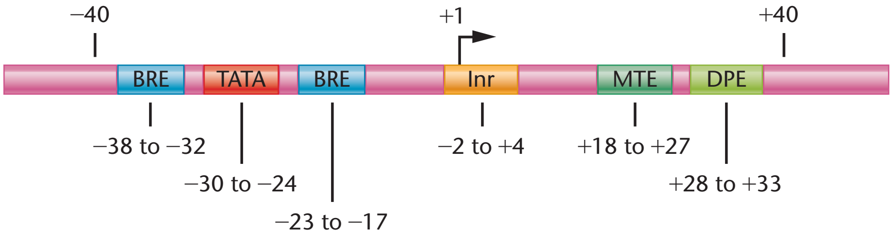


FIGURE 17.5 Core-promoter elements found in focused promoters. Core-promoter elements are usually located between −40 and +40 nucleotides, relative to the transcription start site, indicated as +1. BRE is the TFIIB recognition element, which can be found on one side or other of the TATA box. TATA is the TATA box, Inr is the Initiator element, MTE is the motif ten element, and DPE is the downstream promoter element.

Within core promoters are numerous **core-promoter elements**—short nucleotide sequences that are bound by specific regulatory factors. While it is not yet clear how dispersed promoters specify multiple transcriptional start sites, much more is known about the structure of focused promoters. There is no single universal core promoter; the core-promoter sequence varies between species and between genes within a species. However, focused promoters contain several common DNA sequence elements, which are shown in **Figure 17.5** and described in more detail below.

The Initiator element (**Inr**) encompasses the transcription start site, from approximately nucleotides −2 to +4, relative to the start site. In humans, the Inr consensus sequence is YYAN^{A}/$_{T}$YY (where Y indicates any pyrimidine nucleotide and N indicates any nucleotide). The transcription start site is the first A residue at +1. The **TATA box** is located at approximately −30 relative to the transcription start site and has the consensus sequence TATA^{A}/$_{T}$AAR (where R indicates any purine nucleotide). The TFIIB recognition element (**BRE**) is found at positions either immediately upstream or downstream from the TATA box. The motif ten element (**MTE**) and downstream promoter element (**DPE**) are located downstream of the transcription start site, at approximately +18 to +27 and +28 to +33, respectively.

Versions of the BRE, TATA box, and Inr elements appear to be universal components of all focused promoters; however, the MTEs and DPEs are found in only some of these promoters. Furthermore, this is not a comprehensive list of all known core-promoter elements and there are likely more that have yet to be described. Nonetheless, the important concept here is that the core-promoter elements serve as a platform for the assembly of RNAP II and the general transcription factors, which is a critical step in gene expression and will be discussed in Section 17.4.

In addition to core-promoter elements, many genes also contain proximal-promoter elements located upstream of the TATA box and BRE. Proximal-promoter elements act along with the core-promoter elements to increase the levels of basal transcription. For example, the CAAT box is a common proximal-promoter element. The **CAAT box** has the consensus sequence CAAT or CCAAT and is usually located about 70 to 80 base pairs upstream from the transcription start site. Mutations on either side of this element have no effect on transcription, whereas mutations within the CAAT sequence dramatically lower the rate of transcription. Thus, for genes with a CAAT box, it appears to be required for robust transcription. **Figure 17.6** summarizes the transcriptional effects of mutations in the

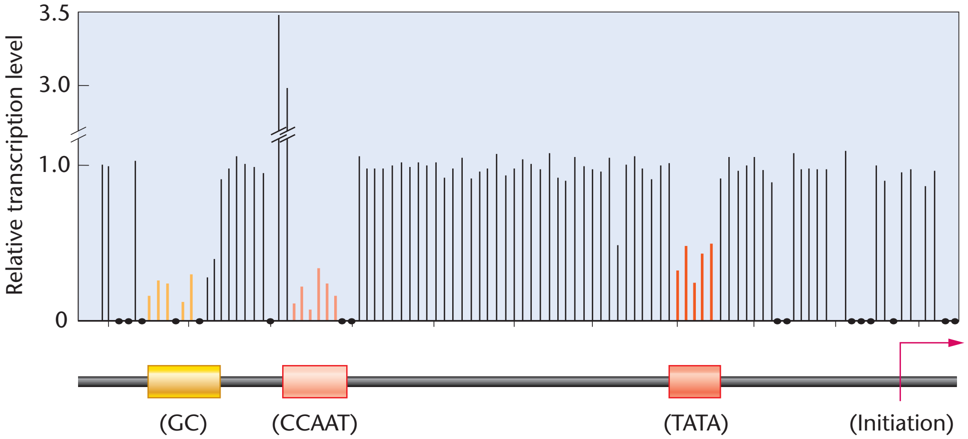


FIGURE 17.6 Summary of the effects on transcription levels of different point mutations in the promoter region of the β-globin gene. Each line represents the level of transcription produced in a separate experiment by a single-nucleotide mutation (relative to wild type) at a particular location. Dots represent nucleotides for which no mutation was tested. Note that mutations within specific elements of the promoter have the greatest effects on the level of transcription.

CAAT box and other promoter elements. The **GC box** is another element often found in promoter regions and has the consensus sequence GGGCGG. It is located, in one or more copies, at about position −110 and is bound by transcription factors.

Enhancers, Insulators, and Silencers

In addition to promoters, transcription of eukaryotic genes is also influenced by DNA sequences called **enhancers.** While promoters are always found immediately upstream of a gene, enhancers can be located on either side of a gene, nearby or at some distance from the gene, or even within the gene. Some studies show that enhancers can be located as far away as a million base pairs from the genes they regulate. Like promoters, they are *cis* regulatory elements because they only serve to regulate genes on the same chromosome. While promoter sequences are essential for minimal or basal-level transcription, enhancers, as their name aptly suggests, increase the rate of transcription. In addition, enhancers often confer time- and tissue-specific gene expression.

Scientists have studied promoters and enhancers by analyzing the effects that specific mutations have on the transcription of **reporter genes** in cultured cells and model organisms. Reporter genes, created by recombinant DNA technology (see Chapter 20), combine a regulatory sequence from a gene of interest with a coding sequence from a gene that confers a phenotype that is easily observed. For example, one popular choice for a reporter gene is the **green fluorescent protein (GFP),** which emits green fluorescence when excited by blue light. In this case, the level of green fluorescence emitted by the cell or tissue is proportional to transcription of the reporter gene. Mutations, deletions, and relocations of promoters and enhancers can thus be tested to determine their effects on transcription by measuring fluorescence intensity. Such studies have revealed two important features that distinguish promoters from enhancers:

1. Whereas promoters must be immediately upstream of the genes they regulate, the position of an enhancer is not critical; it will function the same whether it is upstream, downstream, or within a gene.
2. Whereas promoters are orientation specific, an enhancer can be inverted, relative to the gene it regulates, without a significant effect on its action.

An example of an enhancer located *within* the gene it regulates is the immunoglobulin heavy-chain gene enhancer, which is located in an intron within the gene sequence. This enhancer is required for tissue-specific expression of the immunoglobulin heavy-chain gene. An example of an enhancer located *downstream* of the gene it regulates is the β-globin gene enhancer. In chickens, an enhancer located between the β-globin gene and the ϵ-globin gene works in one direction to control transcription of the ϵ-globin gene during embryonic development and in the opposite direction to regulate expression of the β-globin gene during adult life.

Enhancers are modular and often contain multiple different short DNA sequences. For example, the enhancer of the SV40 virus (which is transcribed inside an infected eukaryotic cell) includes two repeated copies of a 72-bp sequence, located adjacent to each other some 200 bp upstream from a transcriptional start point. Each of the two repeats contains multiple sequence motifs that contribute to enhancing the rate of transcription. If one or the other of these repeats is deleted, there is no effect on transcription; but if both are deleted, *in vivo* transcription is greatly reduced.

If enhancers can regulate genes at a distance and are orientation independent, then what stops an enhancer from coregulating genes on the same chromosome that are transcribed at different times or in different cell types? The extent of an enhancer's reach is limited by boundary elements known as **insulators.** Insulators are generally found between an enhancer and a promoter for a non-target gene. In the aforementioned example, an enhancer located between the β-globin and ϵ-globin genes regulates both genes. There are no insulators between these genes, allowing one enhancer to exert its effect on both genes. However, insulators positioned upstream of the β-globin gene and downstream of the ϵ-globin gene ensure that this particular enhancer does not influence the expression of a gene located outside its region. While the mechanism by which insulators function is not completely clear, evidence suggests that proteins that bind to insulators induce the formation of DNA loops that allow some enhancer–promoter interactions and block others.

Another type of *cis*-acting regulatory element, the **silencer,** acts as a negative regulator of transcription. Silencers, like enhancers, are *cis*-acting short DNA sequence elements that may be located far upstream, downstream, or within the genes they regulate. Another similarity to enhancers is that they also often act in tissue- or temporal-specific ways to control gene expression.

Our understanding of promoters, enhancers, insulators, and silencers has been expanded by data recently reported by the ENCODE project. We will describe the significance of the ENCODE data in Section 17.7.

17.4 Eukaryotic Transcription Initiation Is Regulated by Transcription Factors That Bind to *Cis*-Acting Sites

It is generally accepted that *cis*-acting regulatory sites—including promoters, enhancers, and silencers—influence transcription initiation by acting as binding sites for transcription regulatory proteins, broadly termed **transcription factors.**

In addition to the **general transcription factors (GTFs)** required for the basic process of transcription initiation (see Chapter 13), some transcription factors serve to increase the *levels* of transcription initiation and are known as **activators,** while others *reduce* transcription levels and are known as **repressors.**

The effects of activators and repressors can be modulated for any given cell type, in response to environmental cues, or at the correct time in development. To do this, some transcription factors may be expressed in only certain types of cells, thereby achieving tissue-specific regulation of their target genes. Some transcription factors are expressed in cells only at certain times during development or in response to certain external or internal signals. In some cases, a transcription factor may be present in a cell and may even bind to its appropriate *cis*-acting site but will only become active when modified structurally (for example, by phosphorylation) or by binding to another molecule such as a hormone. These modifications to transcription factors may also be regulated in tissue- or temporal-specific ways. In addition, different transcription factors may compete for binding to the same DNA sequence or to overlapping sequences. Transcription factor concentrations in the cell and the strength with which each factor binds to the DNA will dictate which factor binds, and hence will determine the level of transcription initiation. Finally, the inputs of multiple transcription factors binding to different enhancers and promoter elements are integrated to fine-tune the levels and timing of transcription initiation.

The Human Metallothionein 2A Gene: Multiple *Cis*-Acting Elements and Transcription Factors

The **human metallothionein 2A (*MT2A*) gene** provides an example of how the transcription of one gene can be regulated by the interplay of multiple promoter and enhancer elements and the transcription factors that bind them. The product of the *MT2A* gene is a protein that binds to heavy metals such as zinc and cadmium, thereby protecting cells from the toxic effects of high levels of these metals. The protein is also implicated in protecting cells from the effects of oxidative stress. The *MT2A* gene is expressed at a low or basal level in all cells but is transcribed at high levels when cells are exposed to heavy metals or stress hormones such as glucocorticoids.

The *cis*-acting regulatory elements controlling transcription of the *MT2A* gene include promoter, enhancer, and silencer elements (Figure 17.7). Each element is a short DNA sequence specifically bound by one or more transcription factors.

Several promoter elements, enhancers, and the transcription factors that bind them confer basal levels of transcription of the *MT2A* gene. The core-promoter elements TATA box and Inr are required for transcription initiation. They are bound by RNAP II and several general transcription factors (to be described in Section 17.5). The proximal-promoter element, GC, is bound by the SP1 factor, which is present at all times in most eukaryotic cells and stimulates transcription at low levels. Expression levels are also modulated in response to extracellular growth signals via the activator proteins 1, 2, and 4 (AP1, AP2, and AP4), which are present at various levels in different cell types. AP2 binds an enhancer called the AP2 response element (ARE). AP1 and AP4 bind overlapping sites within the basal element (BLE), which provides some degree of selectivity in how these factors stimulate transcription of *MT2A* when bound to the BLE in different cell types.

High levels of *MT2A* transcription are stimulated by heavy-metal toxicity or stress. The heavy-metal response employs an activator called metal-inducible transcription factor 1 (MTF-1). MTF-1 is normally found in the cytoplasm, but translocates to the nucleus in the presence of heavy metals. Direct metal binding is thought to induce conformational changes that lead to MTF-1's nuclear translocation. In the nucleus, MTF-1 binds several metal response elements (MREs) that enhance *MT2A* transcription.

A different activator and enhancer mediate stress-induced transcription of *MT2A*: the glucocorticoid receptor

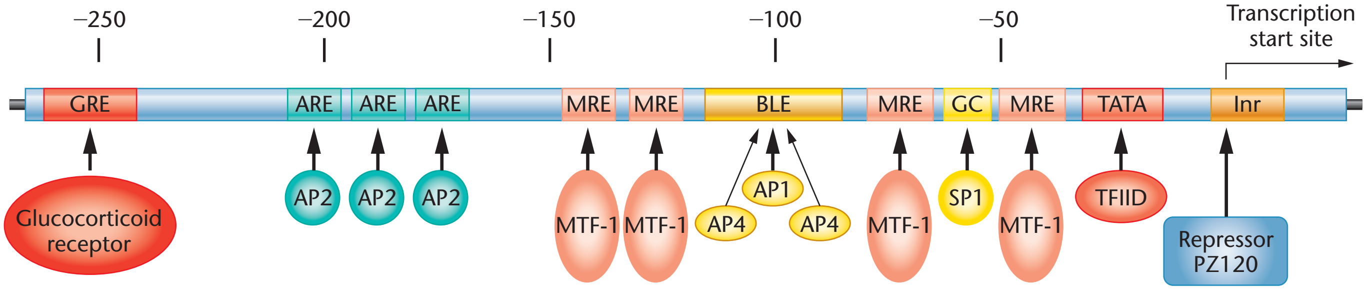


FIGURE 17.7 The human metallothionein 2A gene promoter and enhancer regions, containing multiple *cis*-acting regulatory sites. The transcription factors controlling both basal and induced levels of *MT2A* transcription are indicated below the gene, with arrows pointing to their binding sites.

is an activator that binds to the glucocorticoid response element (GRE). Under stressful conditions, vertebrates secrete a steroid hormone called glucocorticoid. Upon glucocorticoid binding, the glucocorticoid receptor, which is normally located in the cytoplasm, undergoes a conformational change that allows it to enter the nucleus, bind to the GRE, and enhance *MT2A* gene transcription.

In addition to activation, transcription of the *MT2A* gene can be repressed by the actions of the repressor protein PZ120, which binds over the transcription start region.

The presence of multiple regulatory elements and transcription factors that bind to them allows the *MT2A* gene to be transcriptionally activated or repressed in response to subtle changes in both extracellular and intracellular conditions.

Functional Domains of Eukaryotic Transcription Factors

We have described transcription factors as proteins that bind to DNA and activate or repress transcription initiation. These actions are achieved through the presence of two functional domains (clusters of amino acids that carry out a specific function) within each of these proteins. One domain, the DNA-binding domain, binds to specific DNA sequences present in the *cis*-acting regulatory site. The other domain, the **trans-activating domain** or **trans-repressing domain,** activates or represses transcription.

The DNA-binding domains of eukaryotic transcription factors have various characteristic three-dimensional structural motifs. Examples include the helix–turn–helix (HTH), zinc-finger, and basic leucine zipper (bZIP) motifs.

The **helix–turn–helix (HTH) motif,** present in both bacterial and eukaryotic transcription factors, is characterized by a certain geometric conformation rather than a distinctive amino acid sequence. The presence of two adjacent α-helices separated by a "turn" of several amino acids (hence the name of the motif) enables the protein to bind to DNA.

Zinc-finger motifs (such as those present in the MTF-1 transcription factor) are found in a wide range of transcription factors. A typical zinc-finger protein contains clusters of two cysteines and two histidines at repeating intervals. Upon binding zinc atoms, these clusters fold into loops and interact with specific DNA sequences.

The **basic leucine zipper (bZIP) motif** contains a region called a leucine zipper that allows protein–protein dimerization. When two bZIP-containing molecules dimerize, the leucine residues "zip" together. The resulting dimer contains two basic α-helical regions adjacent to the zipper that bind to phosphate residues and specific bases in their target site on DNA.

The *trans*-activating and *trans*-repressing domains of transcription factors bring about their effects by interacting with other transcription factors or RNA polymerase. *Trans*-activating or *trans*-repressing domains may comprise from 30 to 100 amino acids. The amino acid sequences within these domains vary considerably between transcription factors. The ways in which they activate transcription initiation are also varied and are discussed in Section 17.5.

17.5 Activators and Repressors Interact with General Transcription Factors and Affect Chromatin Structure

We have thus far discussed how chromatin remodeling and chromatin modifications are necessary to make *cis*-acting regulatory sequences accessible to binding by transcription factors. In this section we will see how these events come together to regulate the initiation of transcription by facilitating or inhibiting the binding of RNAP II to promoters, and also by regulating RNAP II activity.

Formation of the RNA Polymerase II Transcription Initiation Complex

A critical step in the initiation of transcription is the formation of a **pre-initiation complex (PIC).** The PIC consists of RNAP II and several general transcription factors (GTFs), which assemble onto the promoter in a specific order and provide a platform for RNAP II to recognize transcription start sites and to initiate transcription.

The GTFs that assist RNAP II at a core promoter are called **TFIIA** (*T*ranscription *F*actor for RNAP *IIA*), **TFIIB, TFIID, TFIIE, TFIIF, TFIIH,** and a large multi-subunit complex called **Mediator.** The GTFs and their interactions with the core promoter and RNAP II are outlined in **Figure 17.8** and described below.

The first step in the formation of a PIC (Step 1 in Figure 17.8) is the binding of TFIID to the TATA box of the core promoter. TFIID is a multi-subunit complex that contains **TBP** (*T*ATA *B*inding *P*rotein) and approximately 13 proteins called **TAFs** (*T*BP *A*ssociated *F*actors). In addition to binding TATA, TFIID binds to core-promoter elements such as Inr elements, DPEs, and MTEs. TFIIA interacts with TFIID and assists the binding of TFIID to the core promoter. Once TFIID has made contact with the core promoter, TFIIB binds to BREs on one or the other side of the TATA box (Step 2 in Figure 17.8). Finally, the other GTFs (Mediator, IIF, IIE, and IIH) interact with RNAP II and help recruit it to the promoter (Steps 3 and 4).

The fully formed PIC mediates the unwinding of promoter DNA at the start site and the transition of RNAP II from transcription initiation to elongation. In some higher eukaryotes, RNAP II remains paused on the promoter about

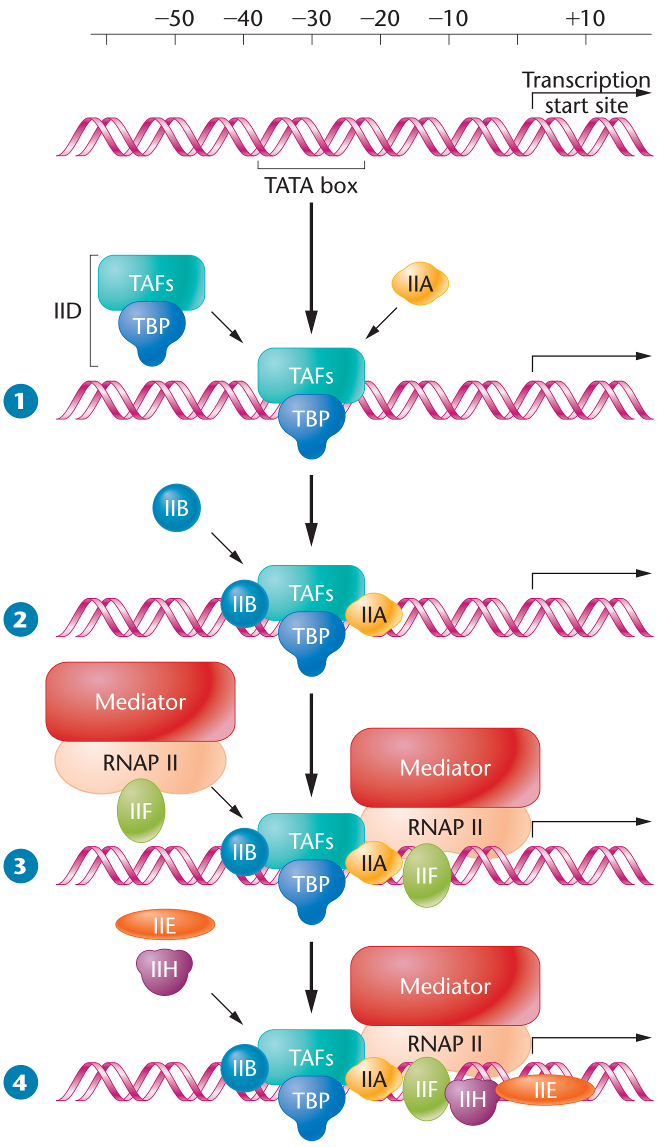


FIGURE 17.8 The assembly of general transcription factors (TFIIA, TFIIB, etc.; abbreviated as IIA, IIB etc.) required for the initiation of transcription by RNAP II.

50 bp downstream of the transcription start site, awaiting signals that release it into transcription elongation. In other higher eukaryotes and in yeast, RNAP II immediately leaves the promoter region and proceeds down the DNA template in an **elongation complex.** Several of the general transcription factors, specifically TFIID, TFIIE, TFIIH, and Mediator, remain on the core promoter to help set up the next PIC.

Mechanisms of Transcription Activation and Repression

Researchers have proposed several models to explain how transcription activators and repressors bring about changes to RNAP II transcription. In most cases, these models involve the formation of DNA loops that bring distant enhancer or silencer elements into close physical contact with the promoter regions of genes that they regulate. Evidence supporting the formation of DNA loops comes from two types of assays: direct visualization and chromosome conformation capture.

Direct visualization assays using fluorescence *in situ* hybridization (FISH) reveal that enhancers and promoters are located in close physical proximity within the nucleus. While useful, this approach is not amenable to high throughput investigations seeking to test interactions between many enhancers and promoters.

More recently, **chromosome conformation capture (3C)** has enabled genome-wide studies of long-range enhancer and promoter interactions. Cells are treated with formaldehyde, which causes crosslinks (covalent bonds) between chromatin regions that are in close physical proximity. DNA sequencing is then used to reveal three-dimensional adjacencies in the genome. Such studies have revealed that enhancer regions and promoter regions, even when separated by millions of base pairs, are in close proximity when promoters are active—presumably through the interactions of proteins bound to each region. In 2015, Eric S. Lander, Erez Lieberman Aiden, and colleagues used chromosome conformation capture to publish a three-dimensional map of the human genome at a high resolution of roughly 1000 base pairs. Their study found ~10,000 DNA loops in the human genome, and ~30 percent of these loops brought together known promoters and enhancers. Keeping in mind that we have not yet annotated all promoters and enhancers in the genome, this study highlights the prevalence of DNA looping as it may relate to transcriptional initiation.

One model of transcription regulation proposes that DNA loops serve to deliver activators, repressors, and GTFs to the vicinity of promoters that must be activated or repressed. In this *recruitment model*, enhancer and silencer elements increase the concentrations of important regulatory proteins at gene promoters. By enhancing the rate of PIC assembly or stability, or by accelerating the release of RNAP II from a promoter, transcription activators bound at enhancers may stimulate the rate of transcription initiation. Direct interactions between activators and repressors with Mediator and TFIID have been documented. Such interactions may serve to close DNA loops between promoters and enhancers. In other cases, proteins called **coactivators** serve as a bridge between activators and promoter-bound factors. Large complexes of activators and coactivators that come together to direct transcriptional activation are called **enhanceosomes** (Figure 17.9). In a similar way, repressors bound at silencer elements may decrease the rate of PIC assembly and the release of RNAP II.

Alternatively, in the *chromatin alteration model*, DNA looping may result in chromatin alterations that either stimulate or repress transcription of target genes. Chromatin remodeling complexes or chromatin modifiers, once

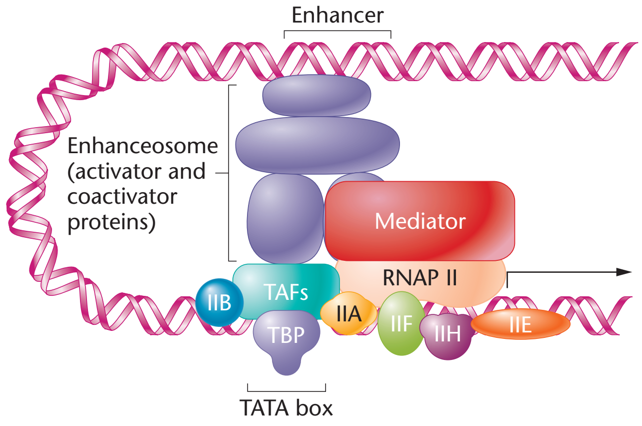


FIGURE 17.9 Formation of a DNA loop allows factors that bind to an enhancer or silencer at a distance from a promoter to interact with general transcription factors in the pre-initiation complex and to regulate the level of transcription.

delivered to the vicinity of a promoter, may open or close the promoter to interactions with GTFs and RNAP II or may inhibit the release of paused RNAP II from pre-initiation complexes. Importantly, this model complements the *recruitment model*; some activators are known to recruit chromatin remodeling complexes or chromatin modifiers, such as histone acetyl transferases (HATs), to DNA loops. Recall from Section 17.2 that histone acetylation by HATs weakens the association between histones and DNA, making DNA within nucleosomes more accessible for transcription. Similarly, some repressors recruit chromatin remodeling complexes or chromatin modifiers, such as histone deacetylases (HDACs), that make DNA less accessible for transcription.

A third model of transcription activation and repression states that enhancer or repressor looping may relocate a target gene to a nuclear region that is favorable or inhibitory to transcription. This *nuclear relocation model* would be consistent with the presence of the transcription factories—regions of the nucleus that contain concentrations of RNAP II and transcription regulatory factors (Section 17.2).

NOW SOLVE THIS

17.2 The hormone estrogen converts the estrogen receptor (ER) protein from an inactive molecule to an active transcription factor. The activated ER binds to *cis*-acting sites that act as enhancers, located near the promoters of a number of genes. In some tissues, the presence of estrogen appears to activate transcription of ER-target genes, whereas in other tissues, it appears to repress transcription of those same genes. Offer an explanation as to how this may occur.

■ **HINT:** *This problem involves an understanding of how transcription enhancers and silencers work. The key to its solution is to consider the many ways that* trans-*acting factors can interact at enhancers to bring about changes in transcription initiation.*

17.6 Gene Regulation in a Model Organism: Transcription of the *GAL* Genes of Yeast

Earlier in the text (see Chapter 16), we considered how *E. coli* controls gene expression to produce proteins needed for using lactose as a carbon source. Recall that in this bacterial system, transcription initiation is controlled at a single promoter because these genes are arranged in an operon. In this section, we will consider how a eukaryote, yeast, efficiently controls gene expression to produce proteins required for the metabolism of galactose. The systems are somewhat analogous in that gene expression is linked to the metabolic needs of the cell. However, the orchestration of *cis*- and *trans*-acting factors to control gene expression is decidedly distinct, reflecting differences in the gene regulatory mechanisms between bacteria and eukaryotes.

The ***GAL* gene system** in yeast served as the initial model system for studying gene regulation in eukaryotes. This system comprises four structural genes (*GAL1, GAL10, GAL2,* and *GAL7*) and three regulatory genes (*GAL4, GAL80*, and *GAL3*). The structural genes encode proteins that transport galactose into the cell and enzymes required to break down galactose for energy. The products of the regulatory genes positively and negatively control the transcription of the structural genes.

Transcription of the *GAL* structural genes is **inducible.** In the absence of galactose, the *GAL* structural genes are not transcribed; in the presence of galactose, transcription begins immediately and the mRNA concentrations for *GAL* structural proteins increase a thousand-fold.

Null mutations in the regulatory gene *GAL4* prevent activation. This underscores the fact that *GAL4* encodes an activator, the Gal4 protein **(Gal4p),** which is required for transcription of the *GAL* structural genes. Gal4p consists of 881 amino acids and functions as a homodimer. It includes a **DNA-binding domain (DBD)** that recognizes and binds specific DNA sequences and an **activation domain (AD)** that activates transcription [Figure 17.10(a)]. Researchers have identified these functional domains by making deletions in the *GAL4* gene and assaying the gene products for their ability to bind DNA and to activate transcription [Figure 17.10(b)].

Through its DBD, Gal4p binds to ~170-bp-long *cis*-acting elements called **UAS$_G$** (*u*pstream *a*ctivation *s*equences of *GAL* genes), which are enhancers for the *GAL* structural genes. Studies have shown that the chromatin structure of UAS$_G$ sites is open; the nucleosomal DNA is only loosely associated with histones. Thus, Gal4p is able to occupy UAS$_G$ sites whether galactose is present or not. Since Gal4p is able to bind UAS$_G$ sites even in the absence

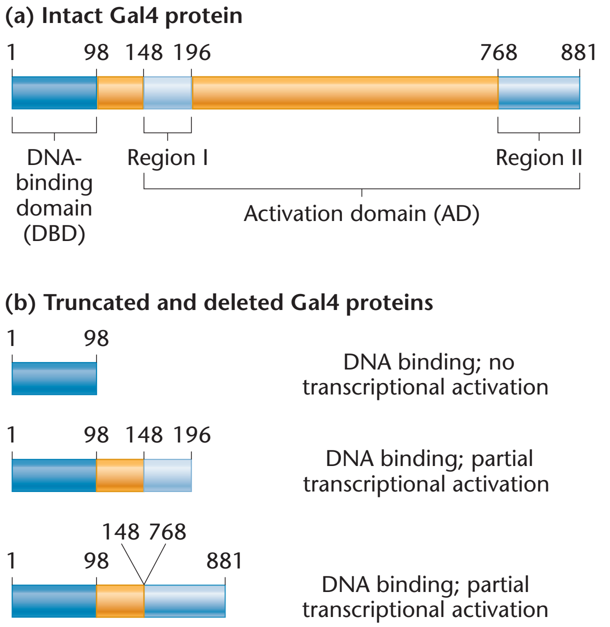

(a) Intact Gal4 protein
1
98
148
196
768
881
DNA-binding domain (DBD)
Region I
Region II
Activation domain (AD)
(b) Truncated and deleted Gal4 proteins
1
98
DNA binding; no transcriptional activation
1
98
148
196
DNA binding; partial transcriptional activation
148
768
1
98
881
DNA binding; partial transcriptional activation

FIGURE 17.10 Structure and function of the Gal4p activator. (a) Gal4p contains a DNA-binding domain, shown in dark blue, and two transcriptional activation regions within the activation domain, shown in light blue. (b) Effects of various deletions on the activity of Gal4p.

of galactose when the *GAL* structural genes are not transcribed, we can conclude that the binding of Gal4p to UAS_G sites is not the sole determinant of transcriptional activation. This involves two other regulatory proteins **Gal80p** and **Gal3p,** encoded by the *GAL80* and *GAL3* genes, respectively.

In the absence of galactose [**Figure 17.11(a)**], Gal80p binds to Gal4p and hides or masks the Gal4p AD. This association inhibits Gal4's ability to activate transcription of the *GAL* structural genes.

In the presence of galactose [**Figure 17.11(b)**], Gal3p binds directly to galactose and undergoes a conformational change that allows it to bind to Gal80p. This interaction relieves Gal4p inhibition leading to the activation of the *GAL* structural genes. The exact nature of the Gal3p—Gal80p interaction is currently under investigation. Recent evidence suggests that Gal80p binds to Gal4p as a homodimer and that Gal3p either destabilizes Gal80p dimers leading to their dissociation, or prevents their association. In this scenario [shown in Figure 17.11(b)], Gal80p monomers may remain bound to Gal4p. Regardless of the exact mechanism, upon galactose binding, Gal3p binds to Gal80p, which enables Gal4p to activate transcription.

Data from many studies suggest that transcriptional activation results from contacts between the Gal4p AD and other proteins. An early step in transcriptional activation by Gal4p is the direct recruitment of a coactivator protein known as SAGA. Following the recruitment of SAGA, the general transcription factor Mediator interacts with Gal4p and enters the complex. Finally, the remaining general transcription factors, as well as RNAP II, are recruited into a pre-initiation complex (PIC) on *GAL* promoters. In the absence of galactose, Gal80p masks the Gal4p AD and thus prevents its association with SAGA. Without the initial recruitment of SAGA, the PIC fails to form on the *GAL* promoters.

Another way in which Gal4p stimulates transcriptional activation is through nucleosome remodeling. Whereas UAS_G sites exhibit open chromatin, *GAL* promoters have nucleosomes positioned at the TATA boxes and Inr elements, making them refractory to PIC formation. Upon induction, Gal4p recruits the nucleosome-remodeling complex SWI/SNF to *GAL* promoters, which leads to the removal of promoter-bound nucleosomes.

Now that we have covered the positive regulation of the *GAL* system by galactose, we will briefly consider how it is negatively regulated in the presence of glucose. When both glucose and galactose are available to the cell, it is more efficient to use glucose and thus conserve the energy required to transcribe and translate the *GAL* structural

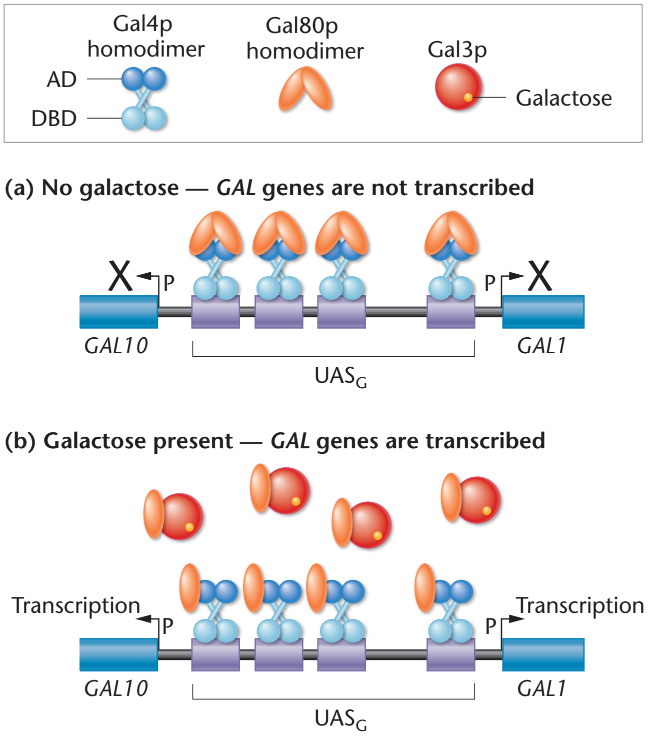


FIGURE 17.11 Model for transcriptional regulation of *GAL* structural genes. Although all four structural genes are regulated by a similar mechanism, only *GAL1* and *GAL10* are shown here. The *GAL10* and *GAL1* gene UAS_G region has four binding sites for Gal4p homodimers. (a) In the absence of galactose, Gal80p homodimers bind Gal4p and mask the activation domain. (b) In the presence of galactose, Gal3p binds to Gal80p and dissociates dimers and/or prevents dimerization leading to exposure of the Gal4p activation domain and transcriptional activation.

genes required for galactose uptake and metabolism. In the presence of glucose, Mig1p, a zinc-finger repressor protein, binds to *cis*-acting silencers for genes in the *GAL* system, such as *GAL1* and *GAL4*. Mig1p binds a corepressor (the Cyc8p/Tup1p complex), which establishes an inaccessible chromatin structure, thus lowering the expression of these genes. However, in the absence of glucose, Mig1p becomes phosphorylated and no longer interacts with Cyc8p/Tup1p, thus leaving the chromatin structure accessible for gene expression.

17.7 ENCODE Data Are Transforming Our Concepts of Eukaryotic Gene Regulation

We conclude this chapter by reviewing a project that has revolutionized our thinking about the human genome and eukaryotic gene regulation. In 2003, the U.S. National Human Genome Research Institute launched the **Encyclopedia of DNA Elements (ENCODE) Project.** The goal of ENCODE is to identify all functional DNA sequences that lie within the 3.1 billion nucleotides of the human genome and to determine how these elements act to regulate gene expression. This massive endeavor, which involves more than 400 scientists from around the world, completed its pilot phase in 2007 and completed its first production project in 2012, which analyzed 1640 datasets from 147 different cell types and produced 30 publications. The data gathered thus far are seriously challenging our views about gene regulation, and additional phases of the ENCODE project are currently under way.

Clearly, the most provocative finding of ENCODE research is that while less than 2 percent of the human genome consists of protein-coding genes, more than 80 percent has a biochemical function! This claim sparked heated discussion within the science community, was widely reported by the media, and challenges us to reconsider what we call "junk DNA"—a term once used to describe all noncoding DNA. Even before the ENCODE project, it was clear that the genome contains many functional regions other than protein-coding sequences such as promoters, enhancers, introns, centromeres, and telomeres. However, the claim that 80 percent of the genome has a biochemical function was met with skepticism from many scientists. As per ENCODE, biochemical function corresponds to DNA sequences that, based on various genome-wide assays, are transcribed, bound by proteins, or associated with specific histone modifications. Does a specific histone modification indicate that the associated sequence has a biochemical function? Does transcription into RNA indicate function, or might some such transcripts reflect "noise" or cellular errors? Ultimately, the term "biochemical function" needs clarification.

While the "biochemical function" debate continues, clearly, ENCODE data include valuable genome annotations of large numbers of enhancers, promoters, sequences encoding RNA transcripts, and sites of specific chromatin modifications.

In this section, we present a brief sampling of ENCODE data as they relate to transcriptional regulation in eukaryotes. We also speculate on how these new data modify our concepts of genetic regulation. More implications of the ENCODE project findings are also presented later in the text (see Chapter 21).

Enhancer and Promoter Elements

By examining chromatin features such as DNase hypersensitivity, which is indicative of open chromatin, and histone modifications, which are characteristic of DNA involved in regulatory-factor binding, ENCODE researchers discovered almost 3 million binding sites for gene-regulatory proteins, including transcription factors that bind to enhancers and promoters. These binding sites include approximately 400,000 enhancer elements and 100,000 promoters. Since the ENCODE project examined 125 different cell types, we have insight into how many of these DNA sites are occupied by transcription factors across all cells types versus those that are unique to a given cell type. Interestingly, about one-third of the sites identified were bound by transcription factors in only one specific cell type, whereas only 3700 sites were bound by transcription factors in all cell types.

As we noted in Section 17.5, activators bound to enhancer elements and repressors bound to silencer elements are thought to work by interacting with GTFs located at promoters. These interactions result in the formation of DNA loops of up to 1 million base pairs in length. ENCODE has examined the interactions between enhancers and promoters, also with surprising results. Although researchers examined only 1 percent of the genome, they identified more than 1000 direct physical interactions between promoters and distant regulatory elements such as enhancers in each cell type examined. Approximately 60 percent of these interactions were specific to only one cell type.

ENCODE data show that enhancers can influence the transcription from several different promoters, and not necessarily the closest one. Only about 7 percent of looping interactions occur between an enhancer and the nearest promoter. Approximately 50 percent of promoters physically interact with more than one distant regulatory element, and some have as many as 20 different interactions.

Similarly, approximately 10 percent of enhancers interact with more than one promoter and sometimes with as many as ten different promoters.

Researchers speculate that these multiple interactions between enhancers and promoters do not all occur simultaneously within the same cell, but most likely reflect the sum of interactions that are occurring within a population of cells that are assayed together in individual experiments. The data suggest that genes and their regulatory elements probably interact in complex and fluid networks. Because enhancer interactions often skip the nearest gene promoter, unknown mechanisms must exist to allow regulatory elements to target the appropriate gene at the right times and places to allow appropriate gene expression.

ENCODE data also indicate that different cell types use different combinations of regulatory elements and binding proteins in order to determine a cell's gene-expression phenotype.

Transcripts and Noncoding RNA

Perhaps the most surprising finding of the ENCODE project is that up to 75 percent of the human genome is transcribed in at least one cell type. This suggests that far more of the genome may be functional than we previously thought—a point that buoys the aforementioned proposition that over 80 percent of the genome has a biochemical function. A total of 128,000 transcript species were identified, and, of these, more than 70,000 arise either from DNA in intergenic regions or from antisense transcription (using the opposite strand of the DNA as a template) within protein-coding genes. More than 7000 small noncoding RNAs were identified, including small nuclear RNAs, small nucleolar RNAs, micro RNAs (see Chapter 18), and tRNAs. Many of these small RNAs are spliced out of pre-mRNAs or long noncoding RNAs.

Another interesting class of RNA identified by the ENCODE research is called **enhancer RNA.** Genomic sequences encoding enhancer RNAs often contain bona fide promoter elements that initiate transcription from the enhancer and extend outward for several kilobases. The most active enhancers in a cell type produce the most enhancer RNA. Overall, the vast majority of transcripts in a cell do not code for proteins and remain within the nucleus.

The functional relevance of this enormous amount of transcription throughout the genome is intriguing but controversial. Some scientists argue that it may simply reflect transcriptional "noise" without functional significance. However, as more and more regulatory roles are being discovered for many of the small noncoding RNAs as well as some long noncoding RNAs (see Chapter 18), it is still possible that functions will be discovered for more of these RNAs.

EVOLVING CONCEPT OF A GENE

Based on the findings of the ENCODE project, we now know that DNA sequences that have previously been described as "junk" DNA, since they do not encode polypeptides, are nonetheless often transcribed into what we call noncoding RNA (ncRNA). Since the function of some of these RNAs is now being determined, we must consider whether the concept of the gene should be expanded to include these DNA sequences. Currently, this is being debated, and given that this is the last "Evolving Concept of a Gene" entry in the text, we are wondering what you think, based on your study of genetics? ■

Many Disease-Associated Genome Variations Affect Regulatory Regions

The identification of mutations associated with inherited disease is important for predicting disease susceptibility, diagnosis, and treatment. For single-gene disorders, such as cystic fibrosis and sickle-cell anemia, disruption of a specific gene is directly correlated to the disease state. However, a myriad of other diseases, such as Parkinson disease, autism, Alzheimer disease, ALS, and cancers, have many genetic variables.

To identify genetic variations associated with disease potential, scientists have turned to genome-wide association studies. A **genome-wide association study (GWAS)** identifies genetic variations that are significantly enriched in the genomes of patients with a particular disease and are not found in those without the disease (see Chapter 22 for enhanced coverage of GWASs).

Many GWASs evaluate single base-pair changes or **single nucleotide polymorphisms (SNPs).** Such studies have yielded insights into the molecular basis of many diseases. However, in many cases, the GWAS-identified SNPs do not directly confer disease susceptibility. Rather the GWAS-identified SNP is located near the causative variant in the genome and, thus, the two are inherited together. This proved to be a limitation to GWAS; however, the analysis of GWAS data in the context of the findings of the ENCODE project has proven extremely powerful.

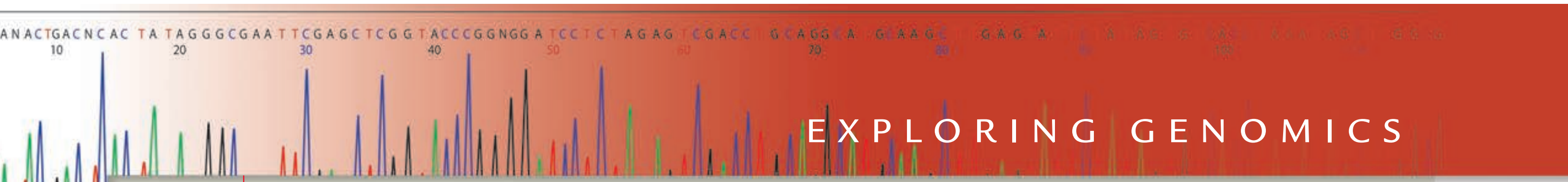
When GWAS information was mapped onto ENCODE data, it was determined that 31 percent of GWAS SNPs are within, or very near a SNP within, an annotated transcription factor binding site. Furthermore, 71 percent are within a DNase-sensitive site, suggesting that such regions have regulatory function. From the mapping data, we conclude that over 90 percent of disease-associated variations in the human genome are located within regulatory regions that likely impact transcription. This alone is remarkable; however, additional insights from the ENCODE project further this finding.

First, many disease-associated SNPs are linked to regulatory regions that confer expression in the specific tissues and developmental stages that correlate to the manifestation of the disease. For example, SNPs associated with multiple sclerosis, an autoimmune disorder, are located within regulatory regions that are DNase sensitive in immune cell types. This supports the hypothesis that an autoimmune disorder may be caused by mutations in regulatory regions that direct active transcription in immune cells. Similarly, SNPs associated with diseases that manifest during fetal development are more likely to be found in fetal-specific DNase-sensitive regions.

Secondly, ENCODE data also link disease-associated SNPs in enhancer regions to the genes that they regulate. As we learned earlier in the chapter, enhancers can exert regulatory effects on genes that are far away. Therefore, it is not trivial to determine which enhancers regulate which genes. However, the ENCODE project annotated thousands of enhancer–promoter interactions in a manner that is cell-type specific. Although these data are far from complete for all enhancers and promoters in all cell types, the data already collected greatly facilitate the identification of target genes disrupted by disease-associated SNPs in regulatory regions.

Finally, when all SNPs associated with a single disease, or group of related diseases, are considered, we find that such SNPs often impact genes that fall into the same regulatory networks. For example, roughly one-quarter of all SNPs associated with a host of inflammatory disorders impact genes within the IRF9 (Interferon regulatory factor 9) pathway. This pathway is known to mediate an immune response and suggests that misregulation of this pathway is associated with many common inflammatory disorders.

We have offered up just a few examples, but as the ENCODE project continues, we can anticipate learning of additional ways in which defects in the regulation of transcription are linked to human disease.



EXPLORING GENOMICS

Tissue-Specific Gene Expression

Mastering Genetics Visit the Study Area: Exploring Genomics

In this chapter, we discussed how gene expression can be regulated in complex ways. One aspect of regulation we considered is the way promoter, enhancer, and silencer sequences can govern transcriptional initiation of genes to allow for tissue-specific gene expression. All cells and tissues of an organism possess the same genome (with some genomic variation as you will learn in Chapter 21), and many genes are expressed in all cell and tissue types. However, muscle cells, blood cells, and all other tissue types express genes that are largely tissue specific (i.e., they have limited or no expression in other tissue types). In this exercise, we return to the National Center for Biotechnology Information (NCBI) site and use the search tool BLAST (Basic Local Alignment Search Tool), which you were introduced to in an earlier "Exploring Genomics" exercise (see Chapter 10). We will use BLAST to learn more about tissue-specific gene-expression patterns.

- **Exercise – Tissue-Specific Gene Expression**

1. Access BLAST from the NCBI Web site at https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. The following are GenBank accession numbers for four different genes that show tissue-specific expression patterns. You will perform your searches on these genes.

   NM_021588.1
   NM_00739.1
   AY260853.1
   NM_004917

3. For each gene, carry out a nucleotide BLAST search using the accession numbers for your sequence query. (Refer to the "Exploring Genomics" feature in Chapter 10 if you need to refresh your memory on BLAST searches.) Because the accession numbers are for nucleotide sequences, be sure to use the "Nucleotide BLAST" program when running your searches. Once you enter "Nucleotide BLAST," under the "Choose Search Set" category, make sure the database is set to "Others (nr, etc.)," so that you are not searching an organism-specific database.
4. Once your BLAST search results appear, look at the top alignment for each gene. Clicking on the link for the top alignment will take you to the page showing the sequence alignment for this gene. To the far right of the page, if you scroll down you will see a section called "Related Information." The "Gene" link provides a report on details related to this gene.

   Some alignments will display a link for "Map Viewer," which will take you to genome mapping information about the gene. The "UniGene" link will show you a UniGene report. For some genes, upon entering UniGene you may need to click a link above the gene name or the gene name itself in order to retrieve a UniGene report. Be sure to explore the “EST Profile” link under the “Gene Expression” category in each UniGene report. EST profiles will show a table of gene-expression patterns in different tissues.

Also explore the “GEO Profiles” link under the “Gene Expression” category of the UniGene reports, when available.

These links will take you to a number of gene-expression studies related to each gene of interest. Explore these resources for each gene, and then answer the following questions:

a. What is the identity of each sequence, based on sequence alignment? How do you know this?

b. What species was each gene cloned from?

c. Which tissue(s) are known to express each gene?

d. Does this gene show regulated expression during different times of development?

e. Which gene shows the most restricted pattern of expression by being expressed in the fewest tissues?

CASE STUDY Risk assessment

Each year in the United States, there are over 230,000 new cases of prostate cancer and almost 28,000 deaths. A 3.8-Mb region on chromosome 8 (8q24), called a gene desert, has very few genes but contains enhancer sequences that potentially confer significant risks for prostate cancer. One particular enhancer allele, which is known to be associated with an elevated risk for prostate cancer, physically interacts with the promoter region of the nearby *MYC* gene and facilitates its upregulation. Overexpression of *MYC*, which encodes a cell-cycle regulatory protein, is observed in multiple types of cancer (see Chapter 24). The risk allele has a frequency of 49 percent in men of European descent and 81 percent in men of African ancestry. Most of the differential *MYC* activity associated with the risk allele occurs during prenatal development, raising the possibility that testing for this allele early in life can be used to identify those in the African-American population who are at very high risk for prostate cancer. However, researchers cannot rule out the possibility that this enhancer causes overexpression of other genes, which may also be involved in prostate cancer.

1. How do enhancers control the expression of genes located some distance away from the enhancer?
2. How can one enhancer control expression of more than one gene?
3. If a screening test for the risk allele is developed, the test could show that the risk allele is not present, and *MYC* activity would be normal. However, someone can receive a negative result from this test but still have a higher than normal risk for prostate cancer from other mutations that contribute to cancer risk. What ethical concerns are there with using a test for cancer susceptibility that is focused on only one risk allele?

For related reading, see Freedman, M. L., et al. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. USA* 103:14068–14073.

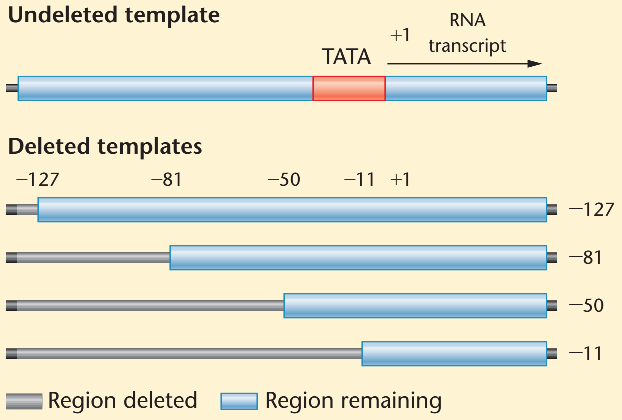
Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. The regulation of gene expression in eukaryotes differs in several ways from that of bacteria.
2. Eukaryotic gene regulation at the chromatin level may involve gene-specific chromatin remodeling, histone modifications, or DNA modifications.
3. Eukaryotic transcription is regulated by *cis*-acting regulatory elements such as promoter, enhancer, and silencer elements.
4. Transcription factors influence transcription by binding to promoters, enhancers, and silencers.
5. Transcription factors regulate transcription by promoting or inhibiting the association of general transcription factors with the core promoter and by modifying chromatin structure.
6. The *GAL* system of yeast serves as a model for studying many important components of transcriptional gene regulation, such as activators, enhancers, and chromatin modifiers.
7. The ENCODE project has analyzed, and continues to analyze, many datasets to catalog DNA sequences in the human genome, such as a genome-wide annotation of regulatory sequences.

INSIGHTS AND SOLUTIONS

1. As a research scientist, you have decided to study transcription regulation of a gene whose DNA has been cloned and sequenced. To begin your study, you obtain the DNA clone of the gene, which includes the gene and at least 1 kb of upstream DNA. You then create a number of subclones of this DNA, containing various deletions in the gene's upstream region. These deletion templates are shown in the figure below.



To test these DNA templates for their ability to direct transcription of the gene, you prepare two different types of *in vitro* transcription systems. The first is a defined system containing purified RNAP II and the purified general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. The second system consists of a crude nuclear extract, which is made by extracting most of the proteins from the nuclei of cultured cells. When you test your two transcription systems using each of your templates, you obtain the following results:

DNA Added	Purified System	Nuclear Extract
Undeleted	+	++++
−127 deletion	+	++++
−81 deletion	+	++++
−50 deletion	+	+
−11 deletion	o	o

+	Low-efficiency transcription
++++	High-efficiency transcription
o	No transcription

(a) Why is there no transcription from the −11 deletion template in both the crude extract and the purified system?

(b) How do the results for the nuclear extract and the purified system differ, for the *undeleted* template? How would you interpret this result?

(c) For each of the various deletion templates, compare the results obtained from both the nuclear extract and the purified systems.

(d) What do these data tell you about the transcription regulation of this gene?

Solution:

(a) The lack of transcription from the −11 template suggests that some essential DNA sequences are missing from this deletion template. As the −50 template does show some transcription in both the crude extract and purified system, it is likely that the essential missing sequences, at least for basal levels of transcription, lie between −50 and −11. As the TATA box is located in this region, its absence in the −11 template may be the reason for the lack of transcription.

(b) The undeleted template containing large amounts of upstream DNA is sufficient to promote high levels of transcription in a nuclear extract, but only low levels in a purified system. These data suggest that something is missing in the purified system, compared with the nuclear extract, and this component is important for high levels of transcription from this promoter. As crude nuclear extracts are not defined in content, it would not be clear from these data what factors in the extract are the essential ones.

(c) Both the −127 and −81 templates act the same way as the undeleted template in both the nuclear extract and the purified system—high levels of transcription in nuclear extracts but low levels in a purified system. In contrast, the −50 template shows only low levels of transcription in both systems. These results indicate that all of the sequences necessary for high levels of transcription in a crude system are located between −81 and −50.

(d) First, these data tell you that general transcription factors alone are not sufficient to specify high efficiencies of transcription from this promoter. The DNA sequence elements through which the general transcription factors work are located within 50 bp of the transcription start site. Second, the data tell you that the promoter for this gene is likely a member of the "focused" class of promoters, with one defined transcription start site and an essential TATA box. Third, high levels of transcription require sequences between −81 and −50 relative to the transcription start site. These sequences interact with some component(s) of crude nuclear extracts.

Problems and Discussion Questions

1. HOW DO WE KNOW? In this chapter, we focused on how eukaryotic genes are regulated at the transcriptional level. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter,

(a) How do we know that promoter and enhancer sequences control the initiation of transcription in eukaryotes?
(b) How do we know that the orientation of promoters relative to the transcription start site is important while enhancers are orientation independent?
(c) How do we know that eukaryotic transcription factors bind to DNA sequences at or near promoter regions?
(d) How do we know that there is an association between disease susceptibility in humans and regulatory DNA sequences?

2. CONCEPT QUESTION Review the Chapter Concepts list on page 393. All of these concepts relate to various ways in which transcription is regulated in eukaryotes. Write a short essay describing how *cis*-acting regulatory elements, activators, and chromatin modifiers are all coordinately involved in regulating transcription initiation.

3. What features of eukaryotes provide additional opportunities for the regulation of gene expression compared to bacteria?

4. Provide a definition of chromatin remodeling, and give two examples of this phenomenon.

5. Describe the organization of the interphase nucleus. Include in your presentation a description of chromosome territories, interchromatin compartments, and transcription factories.

6. A number of experiments have demonstrated that areas of the genome that are transcriptionally inactive are also resistant to DNase I digestion. However, transcriptionally active areas are DNase I sensitive. Describe how DNase I resistance or sensitivity might indicate transcriptional activity.

7. Provide a brief description of two different types of histone modification and how they impact transcription.

8. Present an overview of the manner in which chromatin can be remodeled. Describe the manner in which these remodeling processes influence transcription.

9. Chromatin remodeling by the SWI/SNF complex requires hydrolysis of ATP. What purpose does this serve?

10. Explain how the addition of acetyl groups to histones leads to a weaker association of DNA in nucleosomes.

11. Distinguish between the *cis*-acting regulatory elements referred to as promoters and enhancers.

12. Enhancers can influence the transcription of genes far away on the same chromosome. How are the effects of enhancers restricted so that they do not exert inappropriate transcriptional activation of non-target genes?

13. Describe the manner in which activators and repressors influence the rate of transcription initiation. How might chromatin structure be involved in such regulation?

14. Compare the control of gene regulation in eukaryotes and bacteria at the level of initiation of transcription. How do the regulatory mechanisms work? What are the similarities and differences in these two types of organisms in terms of the specific components of the regulatory mechanisms?

15. Many promoter regions contain CAAT boxes containing consensus sequences CAAT or CCAAT approximately 70 to 80 bases upstream from the transcription start site. How might one determine the influence of CAAT boxes on the transcription rate of a given gene?

16. Research indicates that promoters may fall into one of two classes: *focused* or *dispersed*. How do these classes differ, and which genes tend to be associated with each?

17. Explain the features of the Initiator (Inr) elements, BREs, DPEs, and MTEs of focused promoters.

18. Many transcriptional activators are proteins with a DNA-binding domain (DBD) and an activation domain (AD). Explain how each domain contributes to transcriptional initiation. Would you expect repressors to also have each of these domains?

19. How do the ENCODE data vastly help determine which enhancers regulate which genes?

20. DNA supercoiling, which occurs when coiling tension is generated ahead of the replication fork, is relieved by DNA gyrase. Supercoiling may also be involved in transcription regulation. Researchers discovered that enhancers operating over a long distance (2500 bp) are dependent on DNA supercoiling, while enhancers operating over shorter distances (110 bp) are not so dependent [Liu et al. (2001). *Proc. Natl. Acad. Sci. USA* 98:14,883–14,888]. Using a diagram, suggest a way in which supercoiling may positively influence enhancer activity over long distances.

Extra-Spicy Problems

21. Because the degree of DNA methylation appears to be a relatively reliable genetic marker for some forms of cancer, researchers have explored the possibility of altering DNA methylation as a form of cancer therapy. Initial studies indicate that while hypomethylation suppresses the formation of some tumors, other tumors thrive. Why would one expect different cancers to respond differently to either hypomethylation or hypermethylation therapies?

22. Explain how the following mutations would affect transcription of the yeast *GAL1* gene in the presence of galactose.

(a) A deletion within the *GAL4* gene that removes the region encoding amino acids 1 to 100.
(b) A deletion of the entire *GAL3* gene.
(c) A mutation within the *GAL80* gene that blocks the ability of Gal80 protein to interact with Gal3p.
(d) A deletion of one of the four $\mathrm{UAS_G}$ elements upstream from the *GAL1* gene.
(e) A point mutation in the *GAL1* core promoter that alters the sequence of the TATA box.

23. The interphase nucleus is a highly structured organelle with chromosome territories, interchromatin compartments, and transcription factories. In cultured human cells, researchers have identified approximately 8000 transcription factories per cell, each containing an average of eight tightly associated RNAP II molecules actively transcribing RNA. If each RNAP II molecule is transcribing a different gene, how might such a transcription factory appear? Provide a simple diagram that shows eight different genes being transcribed in a transcription factory and include the promoters, structural genes, and nascent transcripts in your presentation.

24. A particular type of anemia in humans, called β-thalassemia, results from a severe reduction or absence of the normal β-globin chain of hemoglobin. However, the γ-globin chain, normally only expressed during fetal development, can functionally substitute for β-globin. A variety of studies have explored the use of the nucleoside 5-azacytidine for the expression of γ-globin in adult patients with β-thalassemia.

(a) How might 5-azacytidine lead to expression of γ-globin in adult patients?
(b) Explain why this drug may also have some adverse side effects.

25. Regulation of the *lac* operon in *E. coli* (see Chapter 16) and regulation of the *GAL* system in yeast are analogous in that they both serve to adapt cells to growth on different carbon sources. However, the transcriptional changes are accomplished very differently. Consider the conceptual similarities and differences as you address the following.

(a) Compare and contrast the roles of the *lac* operon inducer in bacteria and Gal3p in eukaryotes in the regulation of their respective systems.
(b) Compare and contrast the *cis*-regulatory elements of the *lac* operon and *GAL* gene system.
(c) Compare and contrast how these two systems are negatively regulated such that they are downregulated in the presence of glucose.

26. DNA methylation is commonly associated with a reduction of transcription. The following data come from a study of the impact of the location and extent of DNA methylation on gene activity in eukaryotic cells. A bacterial gene, luciferase, was inserted into plasmids next to eukaryotic promoter fragments. CpG sequences, either within the promoter and coding sequence (transcription unit) or outside of the transcription unit, were methylated to various degrees, *in vitro*. The chimeric plasmids were then introduced into cultured cells, and luciferase activity was assayed. These data compare the degree of expression of luciferase with differences in the location of DNA methylation [Irvine et al. (2002). *Mol. and Cell. Biol.* 22:6689–6696]. What general conclusions can be drawn from these data?

	Size of Methylated Patch (kb)	Number of Methylated CpG Sequences	Relative Luciferase Expression
No methylation	0.0	0	490×
Methylation location: Outside luciferase transcription unit (0–7.6 kb away)	2.0	100	290×
	3.1	102	250×
Methylation location: Inside luciferase transcription unit	1.9	108	80×
	2.4	134	5×
Methylation location: Entire plasmid	12.1	593	2×

27. During an examination of the genomic sequences surrounding the human β-globin gene, you discover a region of DNA that bears sequence resemblance to the glucocorticoid response element (GRE) of the human metallothionein IIA (*hMTIIA*) gene. Describe experiments that you would design to test (1) whether this sequence was necessary for accurate β-globin gene expression and (2) whether this sequence acted in the same way as the *hMTIIA* gene's GRE.

28. Marine stickleback fish have pelvic fins with long spines that provide protection from larger predatory fish. Some stickleback fish were trapped in lakes and have adapted to life in a different environment. Many lake populations of stickleback fish lack pelvic fins. Shapiro et al. (2004) (*Nature* 428:717.723) mapped the mutation associated with the loss of pelvic fins to the *Pitx1* locus, a gene expressed in pelvic fins, the pituitary gland, and the jaw. However, the coding sequence of the *Pitx1* gene is identical in marine and lake stickleback [Chan et al. (2010). *Science* 327:5963,302–305]. Moreover, when the *Pitx1* coding region is deleted, the fish die with defects in the pituitary gland and the jaw, and they lack pelvic fins. Explain how a mutation near, but outside of, the coding region of *Pitx1* may cause a loss of pelvic fins without pleiotropic effects on the pituitary gland and jaw.

29. Although a single activator may bind many enhancers in the genome to control several target genes, in many cases, the enhancers have some sequence conservation but are not all identical. Keeping this in mind, consider the following hypothetical example:

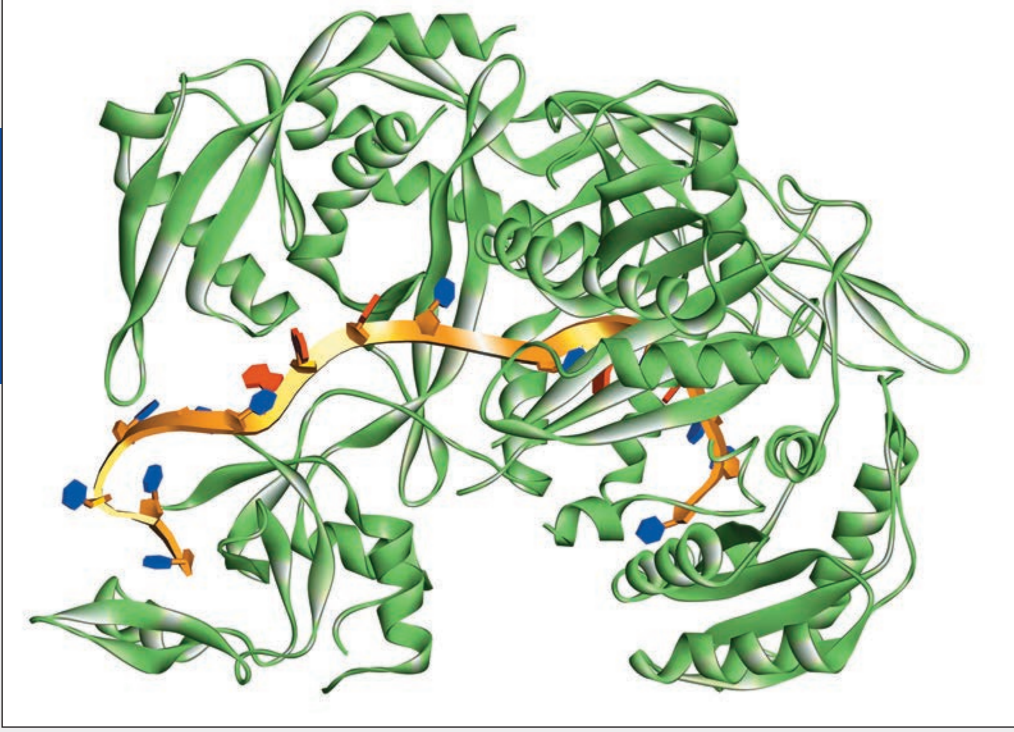
- Undifferentiated cells adopt different fates depending on the concentration of activator protein, Act1.
- A high concentration of Act1 leads to cell fate 1, an intermediate level leads to cell fate 2, and low levels to cell fate 3.
- Research shows that Act1 regulates the expression of three different target genes (A, B, and C) with each having an enhancer recognized by Act1 but a slightly different sequence that alters the affinity of Act1 for the enhancer. Act1 has a high affinity for binding the enhancer for gene A, a low affinity for the gene B enhancer, and an intermediate affinity for the gene C enhancer.

From these data, speculate on how Act1 concentrations can specify different cell fates through these three target genes? Furthermore, which target genes specify which fates?

30. Hereditary spherocytosis (HS) is a disorder characterized by sphere-shaped red blood cells, anemia, and other abnormal traits. Ankyrin-1 (ANK1) is a protein that links membrane proteins to the cytoskeleton. Loss of this activity is associated biochemically to HS. However, Gallagher et al. (2010) (*J. Clin. Invest.* 120:4453–4465) show that HS can also be caused by mutations within a region from −282 to −101 relative to the transcriptional start site, which lead to constitutive transcriptional repression in erythroid cells due to local chromatin condensation. Propose a hypothesis for the function of the −282 to −101 region of the *ANK1* gene.

31. Transcription factors play key roles in the regulation of gene expression, but to do so, they must act within the nucleus. Like most proteins, however, transcription factors are translated in the cytoplasm. To enter the nucleus, transcription factors contain nuclear localization signals, which in some cases can work only when bound to some other molecule such as a steroid hormone. After entering the nucleus, transcription factors must bind to appropriate DNA sites and must interact with other transcription proteins at promoters, enhancers, and silencers. Transcription factors then activate or repress transcription through their activation or repression domains. Many drug therapies target transcription factors. Based on the information provided above, suggest three specific mechanisms through which a successful drug therapy, targeted to a transcription factor, might work.

18



Posttranscriptional Regulation in Eukaryotes

Crystal structure of human Argonaute2 protein interacting with "guide" RNA. Argonaute2 plays an important role in mediating a posttranscriptional RNA-induced silencing pathway.

CHAPTER CONCEPTS

- Following transcription, there are several mechanisms that regulate gene expression, referred to as posttranscriptional regulation.
- Alternative splicing allows for a single gene to encode different protein isoforms with different functions.
- The interaction between *cis*-acting mRNA sequence elements and *trans*-acting RNA-binding proteins regulates mRNA stability, degradation, localization, and translation.
- Noncoding RNAs may regulate gene expression by targeting mRNAs for destruction or translational inhibition.
- Posttranslational modification of proteins can alter their activity or promote their degradation.

The regulation of gene expression in all organisms clearly begins at transcription. A protein-coding gene must first be transcribed before the mRNA can be translated into a functional protein. However, in eukaryotes there are many opportunities for the regulation of gene expression after transcription, referred to as **posttranscriptional regulation.**

As you learned earlier in the text (see Chapter 13), eukaryotic mRNA transcripts are processed by the addition of a 5′ cap, the removal of noncoding introns, and the synthesis of a 3′ poly-A tail. Each of these steps can be regulated to control gene expression. After mature mRNAs are exported to the cytoplasm, they follow different paths: They may be localized to specific regions of the cell; they may be stabilized or degraded; or they may be translated robustly or stored for translation at a later time. Even after translation, protein activity, localization, and stability can be altered through covalent protein modifications. These and other eukaryotic posttranscriptional regulatory mechanisms are summarized in **Figure 18.1**.

Whereas the regulation of transcription depends on transcription factors and DNA regulatory elements (see Chapter 17), many posttranscriptional mechanisms involve RNA-level regulation. Moreover, posttranscriptional regulation is not only centered *on* RNA, but, in some cases, is regulated *by* RNA. Noncoding RNAs play important roles in the regulation of eukaryotic gene expression.

In this chapter, we will explore several important mechanisms and themes of eukaryotic posttranscriptional regulation. As you read on, keep in mind that while scientists have learned a great deal about how genes are regulated at the posttranscriptional level, there are still many unanswered questions for the curious student to ponder.

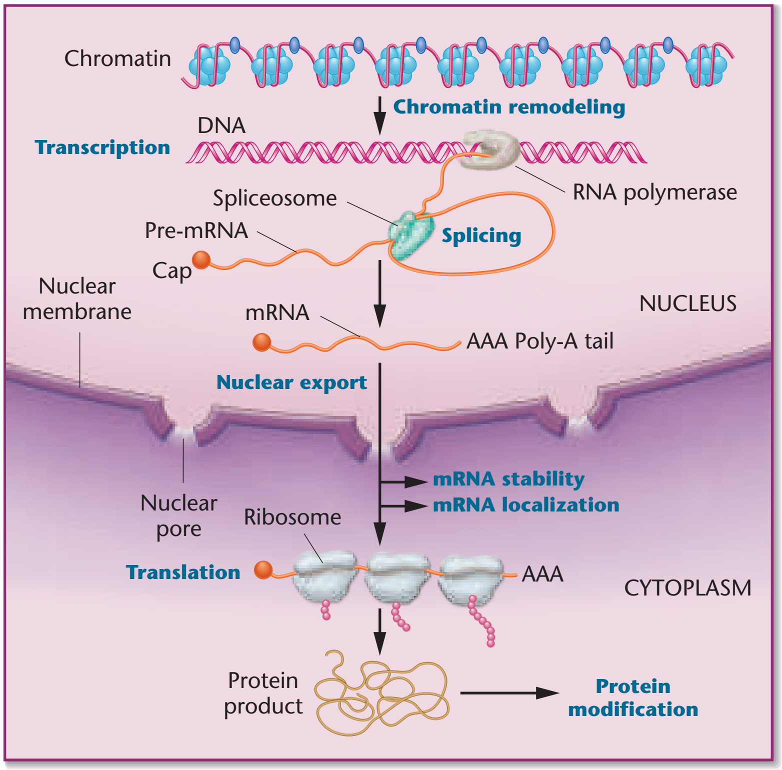


FIGURE 18.1 The regulation of gene expression in eukaryotes. After the initiation of transcription, there are many opportunities for posttranscriptional regulation.

18.1 Regulation of Alternative Splicing Determines Which RNA Spliceforms of a Gene Are Translated

RNA splicing involves the removal of noncoding introns from a pre-mRNA to form a mature mRNA. However, the pre-mRNAs of many eukaryotic genes may be spliced in alternative ways to generate different **spliceforms** that include or omit different exons. This process, known as **alternative splicing,** enables a single gene to encode more than one variant of its protein product. These variants, known as **isoforms,** differ in the amino acids encoded by differentially included or excluded exons. Isoforms of the same gene may have different functions. Even small changes to the amino acid sequence of a protein may alter the active site of an enzyme, modify the DNA-binding specificity of a transcription factor, or change the localization of a protein within the cell. Thus, alternative splicing is important for the regulation of gene expression.

An elegant example of how protein activity can be modulated by alternative splicing is evidenced by the *Drosophila Mhc* gene, which encodes a motor protein responsible for muscle contraction. Different isoforms of this protein are expressed in different types of muscle with slightly different contractile properties. Experiments from the labs of Sanford Bernstein and David Maughan beautifully demonstrated this concept. When an embryo-specific isoform was expressed in flight muscles, it slowed the kinetic properties of the flight muscles and the flies beat their wings at a lower frequency! Thus, in this example we see evidence that alternative splicing regulates gene expression by specifying isoforms with functions that are specific to the cells they are expressed in.

Types of Alternative Splicing

There are many different ways in which a pre-mRNA may be alternatively spliced (Figure 18.2). One example involves **cassette exons**—such exons may be excluded from the mature mRNA by joining the 3′ end of the upstream exon to the 5′ end of the downstream exon. Skipping of cassette exons is the most prevalent type of alternative splicing in animals, accounting for nearly 40 percent of the alternative splicing events.

In slightly over a quarter of the alternative splicing events in animals, splicing occurs at an **alternative splice site** within an exon that may be upstream or downstream of the normally used splice site. While some of these splice events are likely "noise," or errors in splice site selection by the spliceosome (Chapter 13), as we will see below, some instances of alternative splice site usage are important regulatory events.

Intron retention is the most common type of alternative splicing event in plants, fungi, and protozoa, but is rare in mammals. In some cases, introns, which are normally noncoding sequences, are included in the mature mRNAs and are translated, producing novel isoforms. In other cases, intron retention serves to negatively regulate gene expression at the posttranscriptional level; such mRNAs are degraded or are retained in the nucleus.

In rare cases, splicing is co-regulated for a cluster of two or more adjacent exons such that inclusion of one exon leads to the exclusion of the others in the same cluster. The use of these so-called **mutually exclusive exons** allows for swapping of protein domains encoded by different exons.

Pre-mRNAs with different 5′ and 3′ ends may be produced from the same gene due to different transcription initiation and termination sites. Some genes have **alternative promoters,** so they have more than one

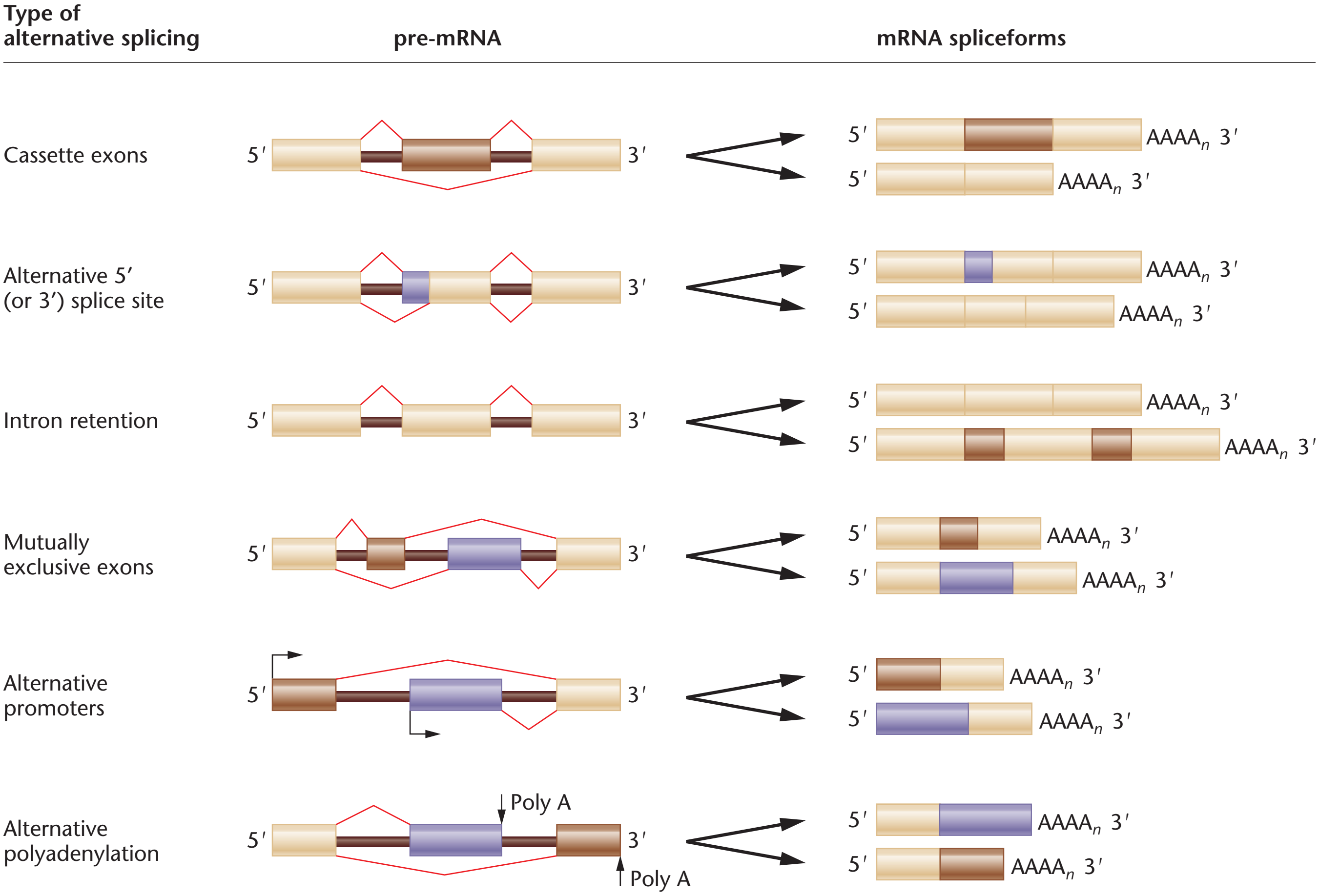


FIGURE 18.2 Different types of alternative splicing events. Exons are indicated by boxes with introns depicted by solid thick lines between them. Alternative splicing is indicated by thin red lines either above or below the pre-mRNA. Transcription start sites (bent arrows) and polyadenylation signals (poly A) are indicated in the alternative promoters and polyadenylation examples.

site where transcription may be initiated. Transcription from alternative promoters produces pre-mRNAs with different 5′ exons, which may be alternatively spliced to downstream exons. Tissue-specific expression of isoforms may result from different transcription factors recognizing different promoters of a gene in different tissues.

Spliceforms with different 3′ ends are produced by **alternative polyadenylation.** The polyadenylation signal (Chapter 13) is a sequence that directs transcriptional termination and addition of a poly-A tail. Thus, when a polyadenylation signal is transcribed, transcription is soon terminated and any downstream exon sequences are omitted. However, when an exon containing a polyadenylation signal is skipped, downstream exons are included and a downstream polyadenylation signal will be used. While alternative polyadenylation may produce spliceforms with different coding sequences, it also specifies different 3′ untranslated regions (UTRs) that are important for other posttranscriptional regulatory events discussed later in this chapter.

Figure 18.3 presents an example of alternative splicing of the pre-mRNA transcribed from the mammalian ***calcitonin/calcitonin gene-related peptide (CT/CGRP)*** gene. In thyroid cells, the *CT/CGRP* pre-mRNA is spliced to produce a mature mRNA containing the first four exons only. In these cells, the polyadenylation signal in exon 4 triggers transcription termination and addition of a poly-A tail. Thus, exons 5 and 6 are omitted. This mRNA is translated and processed into the calcitonin (CT) peptide, a 32-amino-acid peptide hormone that regulates calcium levels in the blood. In neurons, the *CT/CGRP* primary transcript is alternatively spliced to skip exon 4. Since the polyadenylation signal in exon 4 is quickly spliced out during transcription of the pre-mRNA, transcription continues, exons 5 and 6 are included, and the polyadenylation signal of exon 6 is recognized. The *CGRP* mRNA is translated and processed into a 37-amino-acid peptide hormone (CGRP) that stimulates the dilation of blood vessels. Through alternative splicing, the expression of the *CT/CGRP* gene is regulated such that two peptide hormones with different structures and functions are synthesized in different cell types.

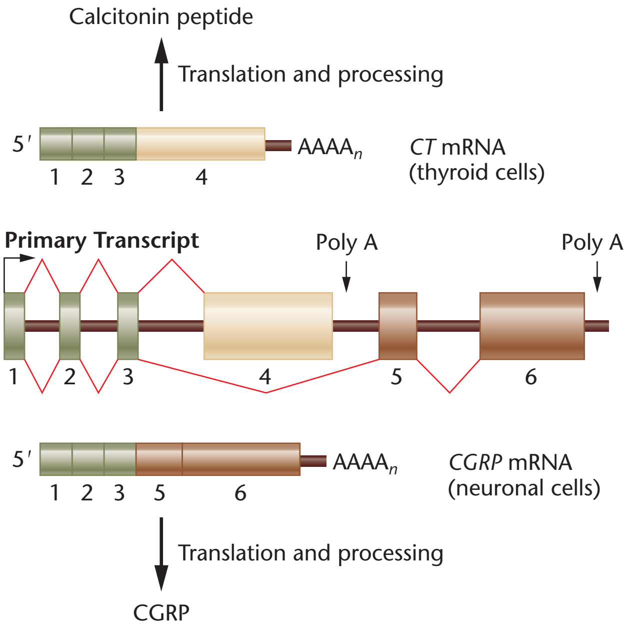


FIGURE 18.3 Alternative splicing of the *CT/CGRP* gene transcript. The primary transcript, which is shown in the middle of the diagram, can be spliced into two different mRNAs, both containing the first three exons but differing in their final exons.

Alternative Splicing and the Proteome

Alternative splicing increases the number of proteins that can be made from each gene. As a result, the number of proteins that an organism can make—its **proteome**—may greatly exceed the number of genes in the genome. Alternative splicing is found in plants, fungi, and animals but is especially common in vertebrates, including humans. Deep sequencing of RNA from human cells suggests that over 95 percent of human multi-exon genes undergo alternative splicing. While not all of these splicing events affect protein-coding sequences, it is clear that alternative splicing contributes greatly to human proteome diversity.

How many different polypeptides can be produced through alternative splicing of the same pre-mRNA? One answer to that question comes from research on the ***Dscam* gene** in *Drosophila melanogaster*. During nervous system development, neurons must accurately connect with each other. Even in *Drosophila*, with only about 250,000 neurons, this is a formidable task. Neurons have cellular processes called axons that form connections with other nerve cells. The *Drosophila Dscam* gene encodes a protein that guides axon growth, ensuring that neurons are correctly wired together. The mature *Dscam* mRNA contains 24 exons; however, the pre-mRNA includes different alternative options for exons 4, 6, 9, and 17 (Figure 18.4). There are 12 alternatives for exon 4; 48 alternatives for exon 6; 33 alternatives for exon 9; and 2 alternatives for exon 17. The number of possible combinations that could be formed in this way suggests that, theoretically, the *Dscam* gene can produce 38,016 different proteins. Although this is an impressive number of isoforms, does the *Drosophila* nervous system require all these alternatives? Recent research suggests that it does.

Each neuron expresses a different subset of Dscam protein isoforms. In addition, *in vitro* studies show that each Dscam isoform binds to the same isoform, but not to others. Even a small change in amino acid sequence reduces or eliminates the binding between two Dscam molecules. *In vivo* studies show that cells expressing the same Dscam isoforms interact with each other. Therefore, it appears that the diversity of Dscam isoforms provides a molecular identity tag for each neuron, helping guide axons to the correct target and preventing miswiring of the nervous system.

The *Drosophila* genome contains about 14,000 protein-coding genes, but the *Dscam* gene alone encodes 2.5 times that many proteins. Because alternative splicing is far more common in vertebrates, the suite of proteins that can be produced from the human genome may be astronomically high. A large-scale mass spectrometry study of the human proteome found that the ~20,000 protein-coding genes in the human genome can produce at least 290,000 different proteins. See Chapter 21 for additional information on proteomes.

Regulation of Alternative Splicing

For pre-mRNAs that are alternatively spliced, how are specific splicing patterns selected? How does the spliceosome select one splice site instead of another? We know that this process is highly regulated, with some spliceforms only present in some cell types or under certain conditions. The regulation of this process involves familiar themes.

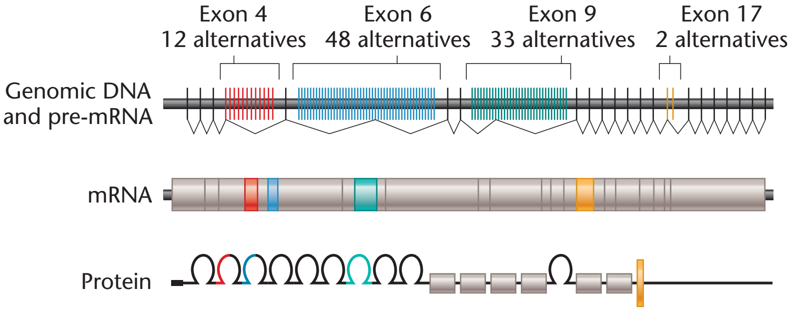


FIGURE 18.4 Alternative splicing of the *Drosophila Dscam* gene mRNA. Organization of the *Dscam* gene and the pre-mRNA. The *Dscam* gene encodes a protein that guides axon growth during development. Each mRNA will contain one of the 12 possible exons for exon 4 (red), one of the 48 possible exons for exon 6 (blue), one of the 33 possible exons for exon 9 (green), and one of the 2 possible exons for exon 17 (yellow).

Cis-acting sequences that regulate alternative splicing are known as **splicing enhancers** and **splicing silencers,** which promote or inhibit, respectively, the splicing of nearby splice sites. A class of proteins known as **SR proteins,** which contain repeats of serine (S) and arginine (R), bind to splicing enhancers and activate splicing by recruiting spliceosome components. **Heterogeneous nuclear ribonucleoproteins (hnRNPs)** are a class of proteins that bind splicing silencers and inhibit splicing. In some cases, hnRNP binding prevents spliceosome assembly at a nearby splice site. Alternatively, hnRNPs may exclude exons by binding to splicing silencers on either side of an exon and then binding to each other.

Many **RNA-binding proteins (RBPs),** which are a class of proteins that bind to specific RNA sequences or RNA secondary structures, are also involved in the regulation of alternative splicing. Since RBPs often exhibit tissue-specific expression, they are important regulators of tissue-specific alternative splicing. RBPs may act by binding and hiding splice sites to promote the use of alternative sites. Other RBPs bind to splicing enhancers or silencers to prevent the binding of SR proteins or hnRNPs. Other RBPs directly interact with the splicing machinery.

Sex Determination in *Drosophila*: A Model for Regulation of Alternative Splicing

As outlined earlier in the text (see Chapter 7), sex in *Drosophila* is determined by the ratio of X chromosomes to sets of autosomes (X : A). When the X : A ratio is 0.5 (X : 2A), males are produced (in *Drosophila*, the Y chromosome is not male-determining and thus is ignored in sex-determination ratios). When the X : A ratio is 1.0 (2X : 2A), females are produced. Chromosomal ratios are interpreted by a small number of genes that initiate a cascade of alternative splicing events, which produce proteins that direct male or female development. Three major genes in this pathway are *Sex lethal* (*Sxl*), *transformer* (*tra*), and *doublesex* (*dsx*). We will review some of the key steps in this process.

The regulatory gene at the beginning of this cascade (**Figure 18.5**) is the gene ***Sex lethal (Sxl),*** which encodes an RNA-binding protein. In females, transcription factors encoded by genes on the X chromosome activate transcription of the *Sxl* gene. In males, the lower concentration of these transcription factors is not sufficient to activate transcription of *Sxl.* As a result, the SXL protein is expressed only in female embryos. The presence (in females) or absence (in males) of SXL protein begins a cascade of male- or female-specific pre-mRNA splicing events. In the presence of SXL, the female splicing patterns override the default male splicing patterns.

One of the targets of SXL protein is the pre-mRNA encoded by the ***transformer (tra)*** gene, which is transcribed in both males and females. In the absence of SXL protein, as in males, the *tra* pre-mRNA is spliced such that a UAG stop codon near the 5′ end of exon 2 remains in the mature mRNA. Translation of this male-specific mRNA leads to a truncated and nonfunctional protein. However, in females, SXL protein binds to a splicing silencer and promotes splicing to an alternative 3′ splice site downstream of the early stop codon. Translation of the female-specific spliceform of *tra* leads to a full-length and functional SR protein (TRA), which regulates the alternative splicing of several targets.

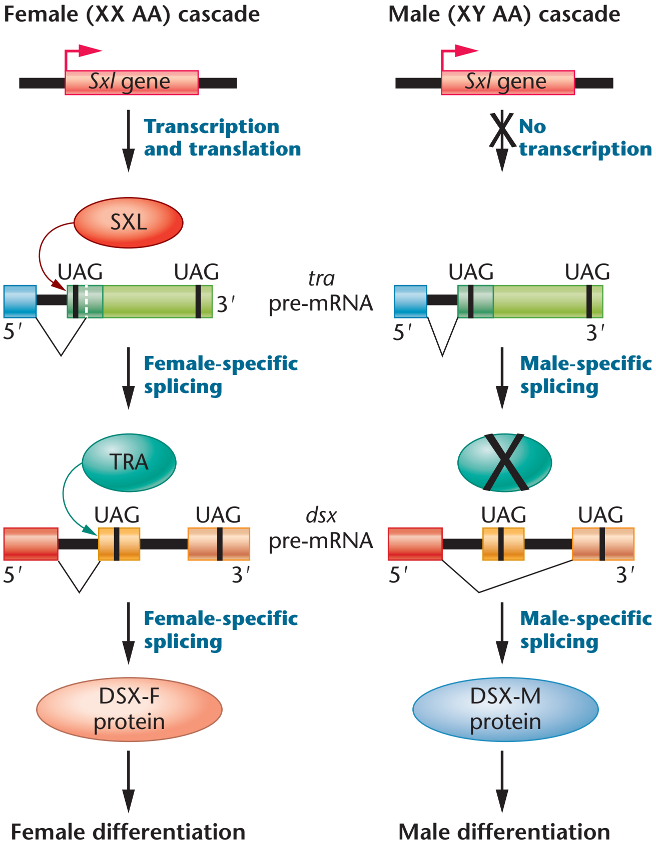


FIGURE 18.5 Regulation of pre-mRNA splicing that determines male and female sexual development in *Drosophila*. The ratio of X chromosomes to autosomes (AA) leads to transcription of the *Sxl* gene in females. The presence of SXL protein begins a cascade of pre-mRNA splicing events that culminate in female-specific gene expression and production of the DSX-F transcription factor. In the absence of SXL protein, a male-specific pattern of pre-mRNA splicing results in male-specific patterns of gene expression induced by the DSX-M protein.

The next gene in the cascade, ***doublesex (dsx),*** is a critical control point in the development of the sexual phenotype. It produces a functional mRNA and protein in both females and males. However, there are male- and female-specific spliceforms of *dsx*, which are translated into different DSX isoforms with distinct functions. In females, the TRA SR protein binds to a splicing enhancer on the *dsx* pre-mRNA and directs splicing in a female-specific pattern making use of an early stop codon and polyadenylation signal. In males, no functional TRA protein is present, and splicing of the *dsx* pre-mRNA involves exon skipping and results in a male-specific mRNA and protein. The female DSX protein (DSX-F) and the male protein (DSX-M) are both transcription factors, but they act in different ways. DSX-F represses the transcription of genes whose products control male sexual development, whereas DSX-M activates the transcription of genes whose products control male sexual development. In addition, DSX-M represses the transcription of genes that control female sexual development.

In sum, different alternative splicing events in males and females regulate sex-specific gene expression patterns.

NOW SOLVE THIS

18.1 Some mutations in the *tra* gene of *Drosophila* cause XX individuals to appear as males. In contrast, other mutations in *tra* cause XY flies to appear as females. Which of these phenotypes would you expect if a mutation in *tra* resulted in a null allele? Which would you expect if the mutation produced a constitutively active *tra* gene product?

■ **HINT:** *This problem involves an understanding of the* Sxl *regulatory pathway in* Drosophila*. The key to its solution is to review the functions of the normal TRA protein in* Drosophila *males and females.*

Alternative Splicing and Human Diseases

Since alternative splicing is an important mechanism for the regulation of gene expression, it is not surprising that defects in alternative splicing are associated with human diseases. Genetic disorders caused by mutations that disrupt RNA splicing are known as **spliceopathies.**

Myotonic dystrophy (DM) is an autosomal dominant disorder that afflicts 1 in 8000 individuals. DM patients exhibit myotonia (inability to relax muscles), muscle wasting, insulin resistance, cataracts, intellectual disability, and cardiac muscle problems. Studies have shown that several of these symptoms are caused by widespread alternative splicing defects in muscle cells and neurons.

There are two forms of DM (DM1 and DM2), which are caused by mutations in different genes, but with similar outcomes that lead to splicing defects. DM1 is caused by expansion of a CTG repeat in the 3′ UTR of the *DMPK* gene. Unaffected individuals have 5–35 copies of the CTG repeat, whereas DM1 patients have 150–2000 copies. The severity of the symptoms is directly related to the number of repeats. DM2 is caused by an expansion of a CCTG repeat sequence within the first intron of the *CNBP* gene (also known as *ZNF9*). Unaffected individuals have 11–26 repeats, while DM2 patients have over 11,000 copies of this repeat; the severity of symptoms is not related to the number of repeats.

Interestingly, DM is not caused by defects in the proteins encoded by *DMPK* and *CNBP*. Rather, repeat-containing RNAs accumulate in the nucleus, instead of being exported to the cytoplasm, and are bound by proteins that regulate alternative splicing. In this way, these RNAs sequester splicing regulators and prevent them from regulating many RNAs that encode proteins important for muscle and neuron function. Strategies to degrade the repeat-containing RNAs, or to block the binding of the splicing regulators to the RNAs, are currently being researched for therapeutic purposes.

Another example of a disorder linked to defective splicing is **spinomuscular atrophy (SMA),** a recessive disorder characterized by the loss of motor neurons that control muscle movement. This leads to muscle atrophy, and eventually patients are unable to breathe. SMA is the most common genetic cause of infant death. SMA is caused by mutations in the *SMN1* (*survival of motor neurons 1*) gene, which encodes SMN, an RNA-binding protein implicated in assembly of the spliceosome. Loss of SMN leads to widespread splicing defects in motor neurons suggesting that SMN is important for proper splicing.

Interestingly, a neighboring gene, called *SMN2*, is also able to produce functional SMN protein. However, 80–90 percent of the time *SMN2* pre-mRNA is improperly spliced, omitting exon 7. This results in a frameshift and an early stop codon, which, when translated, results in a truncated and nonfunctional protein. In theory, if the proportion of spliceforms containing exon 7 could be upregulated, it would result in a substantial increase of functional SMN protein and compensate for a mutation in *SMN1*. In December of 2016, the FDA approved Sprinraza™, a drug that does just that. Sprinraza™, is an **antisense oligonucleotide (ASO)**—a synthetic nucleotide sequence that is complementary, or antisense, to a splicing silencer that promotes exon 7 exclusion. Through complementary base pairing, the ASO blocks the splicing silencer and the exon is no longer spliced out. The results are promising. Clinical trials were stopped early because the benefits to the SMA patients—children—were so obvious that it was deemed unethical to continue giving the control group a placebo. The success of this drug is exciting for SMA families and may reinvigorate other ASO drug investigations. For additional information on the treatment of human diseases using ASO technology, see the earlier Genetics, Ethics, and Society essay on this topic (see Chapter 13.)

18.2 Gene Expression Is Regulated by mRNA Stability and Degradation

The **steady-state level** of an mRNA—meaning the total amount at any one point in time—is a function of the rate at which the gene is transcribed and the rate at which the mRNA is degraded. The steady-state level determines the amount of mRNA that is available for translation. mRNA stability can vary widely between different mRNAs, lasting a few minutes to several days, and can be regulated in response to the needs of the cell. Thus, the molecular mechanisms that control mRNA stability and degradation play important roles in the regulation of gene expression.

Mechanisms of mRNA Decay

RNA is susceptible to degradation, or decay, by **exoribonucleases**—enzymes that degrade RNA via the removal of terminal nucleotides. However, two features of mRNAs provide protection against exoribonucleases: a 7-methylguanosine (m^7G) cap at the 5′ end and a poly-A tail at the 3′ end. Maintenance or removal of the cap and poly-A tail are thus critical steps in determining the stability or decay of an mRNA.

Most eukaryotic mRNAs are degraded by **deadenylation-dependent decay** (Figure 18.6). This process is initiated by **deadenylases,** which are enzymes that shorten the poly-A tail. A newly synthesized mRNA has a poly-A tail that is about 200 nucleotides long. However, if deadenylation shortens it to less than ~30 nucleotides, the mRNA will be degraded. In some cases, an exoribonuclease-containing **exosome complex** destroys the mRNA in a 3′ to 5′ manner. In other cases, the shortened poly-A tail leads to the recruitment of **decapping enzymes,** which remove the 5′ cap and allow a specific exoribonuclease, **XRN1,** to destroy the mRNA in a 5′ to 3′ direction.

More rarely, mRNAs may meet their demise through **deadenylation-independent decay.** This pathway begins at the 5′ cap rather than the 3′ poly-A tail. Similar to deadenylation-dependent decay, decapping enzymes are recruited to remove the cap, and then the XRN1 exoribonuclease digests the mRNA in the 5′ to 3′ direction. In addition, in this pathway, mRNAs may also be cleaved internally by **endoribonucleases.** Following endoribonucleolytic attack, newly formed 5′ and 3′ ends are unprotected and subject to exoribonuclease digestion. This is an important step in a noncoding-RNA–mediated mRNA degradation pathway that we will discuss later in this chapter.

The primary purpose of mRNA decay is to stop translation. Thus, it is no coincidence that mRNA decay pathways are often initiated by removal of the cap or poly-A tail. These events not only make mRNAs susceptible to exoribonuclease digestion, they also remove the cap-binding translation initiation factor eIF4E and the poly-A binding proteins, both of which are critical for translation initiation (see Chapter 14).

mRNAs that are not actively being translated often accumulate in cytoplasmic complexes known as **processing bodies (P bodies).** While the exact composition of P bodies is unknown and is variable in different cell types, we do

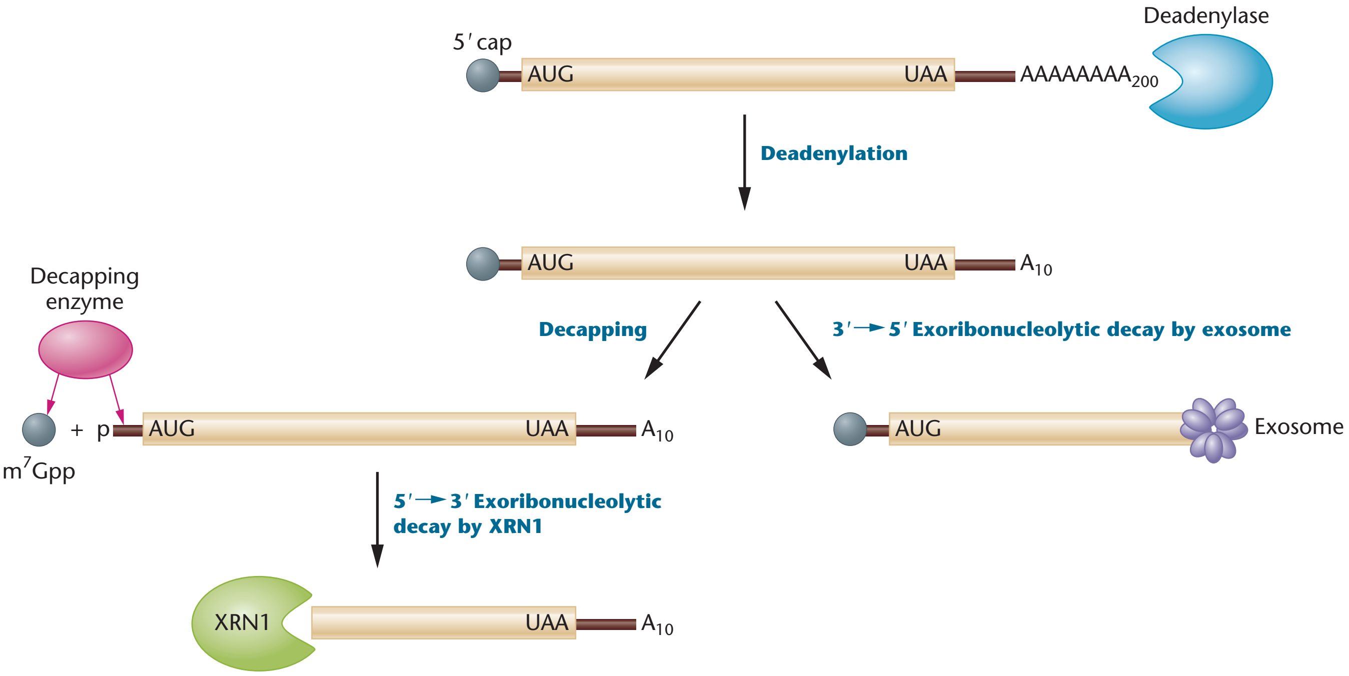


FIGURE 18.6 Deadenylation-dependent mRNA decay.

know that P bodies contain decapping enzymes and exoribonucleases like XRN1. Mounting evidence suggests that P bodies are localized centers for mRNA decay. However, there is also evidence that some mRNAs that end up in P bodies are not destroyed and are able to exit the P bodies and be translated. This suggests that some mRNAs are stored in P bodies for translation at a later time.

Regulation of mRNA Stability and Degradation

While factors such as the decapping enzymes and deadenylases play important roles in initiating mRNA decay, what determines which mRNAs are degraded, and which are not? mRNA decay and stability are controlled through *cis*-acting sequence elements in the mRNAs and the *trans*-acting proteins that bind them.

One well-studied *cis*-acting sequence element that regulates mRNA stability is the **adenosine–uridine rich element (ARE)**—a stretch of A and U nucleotides often found in 3′ UTRs of mRNAs. Roughly 10 percent of mammalian mRNAs contain AREs. ARE binding by RBPs stimulates mRNA stability or degradation. Some RBPs, such as tristetraprolin (TTP), promote mRNA degradation by recruiting decay machinery such as deadenylases and the exosome. Other RBPs, such as HuR, block the recruitment of decay machinery, thus stabilizing the mRNAs.

ARE-mediated mRNA regulation has clinical relevance. Many ARE-containing mRNAs encode proteins that promote cellular proliferation, too much of which may lead to cancer. Interestingly, TTP is downregulated in many cancers. Since TTP promotes the decay of mRNAs with AREs, cancer cells have higher levels of these mRNAs. For example, the ARE-containing *E2F1* mRNA encodes a transcription factor that promotes cellular proliferation. In cancer cells with reduced TTP levels, *E2F1* activity is elevated. Importantly, when TTP was upregulated in cultured cancer cells, *E2F1* mRNA levels dropped and cellular proliferation was significantly reduced. Current research is exploring drug-induced activation of TTP as a cancer therapy.

In addition to AREs, other sequence elements and RBPs have been identified that regulate mRNA stability and decay. However, we are still far from understanding the complex interactions of mRNAs and RBPs that determine mRNA fate.

mRNA Surveillance and Nonsense-Mediated Decay

Aberrant mRNAs can lead to nonfunctional proteins if translated. Eukaryotic cells have evolved several ways to eliminate these potentially harmful mRNAs. For example, mRNAs that lack poly-A tails or are improperly spliced may be retained in the nucleus to allow more time for processing or may be degraded by exoribonucleases.

mRNAs with a premature stop codon trigger an **mRNA surveillance** response. Premature stop codons may result from a nonsense mutation in the gene or be due to an RNA polymerase error during transcription. Translation of an mRNA with a premature stop codon would lead to a truncated and nonfunctional protein, which is a waste of cellular energy and resources and could possibly have toxic effects. However, **nonsense-mediated decay (NMD),** the most thoroughly studied mRNA surveillance pathway, efficiently eliminates mRNAs with premature stop codons.

How does NMD work? Research suggests that recognition of premature stop codons often occurs when translation terminates too far from the poly-A tail, upstream of an exon–exon junction, or upstream of other specific sequences. Once identified, mRNAs with premature stop codons are quickly degraded. In yeast and mammalian cells, decay is most often initiated by decapping enzymes or deadenylases, followed by rapid exoribonuclease digestion. In other species, such as *Drosophila*, NMD involves endoribonuclease attack near the premature stop codon and subsequent exoribonuclease digestion.

18.3 Noncoding RNAs Play Diverse Roles in Posttranscriptional Regulation

In addition to mRNAs that encode proteins, there are several types of **noncoding RNAs (ncRNAs)** that serve a variety of functions in the eukaryotic cell. You should already be familiar with rRNAs and tRNAs, which play important roles in translation (Chapter 14), and snRNAs, which mediate RNA splicing (Chapter 13). Other ncRNAs serve as efficient and specific regulators of posttranscriptional gene expression.

Research on regulatory ncRNAs in eukaryotes began in the 1990s, but has exploded more recently. We now know that ncRNAs are important regulators of gene expression in countless biological contexts, and their dysfunction has been implicated in human disease. RNA-mediated gene regulatory mechanisms have been exploited for the investigation of gene function, biotechnology, and the treatment of diseases. Despite this “RNA revolution,” we still do not have answers to some basic questions. How many regulatory ncRNAs are encoded by eukaryotic genomes? What is the complete repertoire of functions for ncRNAs? In this section we will discuss what *is* known about noncoding RNAs.

The Discovery of RNA Interference and microRNAs

Important scientific breakthroughs are often preceded by serendipitous discovery, as was the case for the discovery of **RNA interference (RNAi)**—a mechanism by which ncRNA molecules guide the posttranscriptional silencing of mRNAs in a sequence-specific manner. Experiments in the early 1990s with petunias, bread mold, and worms revealed that exogenously introduced nucleic acids led to the posttranscriptional silencing of endogenous genes with the same sequence. However, it was unclear why or how this was happening.

Andrew Fire and Craig Mello dedicated their research to the investigation of this phenomenon and were awarded the Nobel Prize in Physiology or Medicine in 2006 for their work. They injected roundworms (*Caenorhabditis elegans*) with either single-stranded or double-stranded RNA molecules—both containing sequences complementary to the mRNA of the *unc-22* gene. Although they expected that the single-stranded antisense RNA molecules would suppress *unc-22* gene expression by binding to the endogenous sense mRNA, they were surprised to discover that the injection of double-stranded *unc-22* RNA was 10- to 100-fold more powerful in repressing *unc-22* mRNA. They studied the phenomenon further and published their results in 1998. They reported that the presence of double-stranded RNA leads to a "potent and specific" degradation of complementary mRNA. Only a few molecules of double-stranded RNA are needed to bring about the degradation of large amounts of mRNA. These findings opened up an entirely new and surprising branch of molecular biology, with far-reaching implications for practical applications.

Around the same time that RNAi was discovered, two other research teams, also working with *C. elegans*, made another discovery related to RNA-mediated posttranscriptional regulation. Victor Ambros's lab showed that a mutation in a gene that controls developmental timing in the worm did not encode a protein. Rather, it encodes a small RNA complementary in sequence to an mRNA encoding a protein that also regulates developmental timing. Ambros's lab had discovered a type of regulatory ncRNA that we now refer to as **microRNA (miRNA).** Gary Ruvkun's lab showed that complementary base pairing between the miRNA and the mRNA leads to a translational downregulation of the worm mRNA. In addition, David Baulcombe's lab showed that miRNAs also function in plants. For their discoveries, Ambros, Ruvkun, and Baulcombe were awarded the 2008 Albert Lasker Award for Basic Medical Research—a high honor that often precedes a Nobel Prize.

The most important concept that we have learned from these seminal discoveries is that ncRNAs can associate with mRNAs through complementary base pairing, and regulate them via destruction or translation inhibition. Next, we will explore how this happens.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

MicroRNAs Regulate Ovulation in Female Mice

In this chapter you learned about microRNAs (miRNAs) and their role in posttranscriptional gene silencing. Using the same methods for knocking out protein-coding genes, scientists have created knockouts to study the functions of miRNAs. We will discuss in detail how knockout animals are generated later in the text (see Chapter 20). Through these and other studies we now know that miRNAs have important roles during development and adult life.

Here we consider a recent study of the role of two miRNAs, miR-200b and miR-429, in reproduction. Both are in the miR-200 family of related sequences and differ from one another by just a single nucleotide, suggesting that they may regulate the same target mRNAs.

miR-200b had been detected in mouse testes; its expression peaked at about two weeks of age and continued through adulthood. Because of this finding, scientists hypothesized that miR-200b might play a role in male fertility. To test this hypothesis, a line of miR-200b/miR-429 double knockout (DKO) mice was generated.

Results:

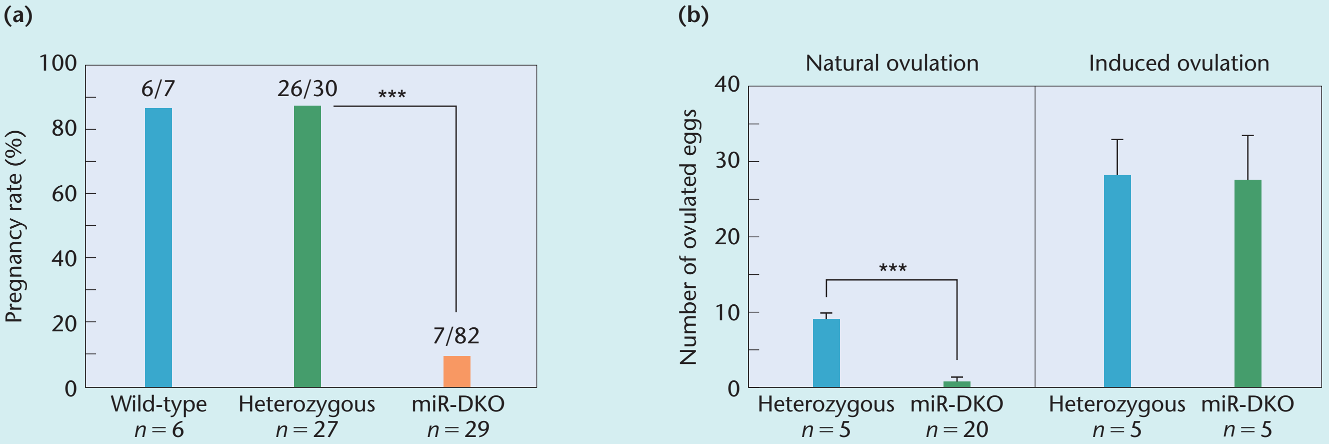
Researchers were surprised that no abnormalities were found in the testes and there were no fertility defects in DKO male mice. In contrast, miR-DKO female mice showed greatly reduced fertility. The pregnancy rate of miR-DKO females was at best 9 percent compared to 85 percent in wild-type females (see panel a). This fertility difference persisted even after three months of attempted mating with male mice.

A substantial decrease in serum levels of luteinizing hormone (LH), a hormone essential for ovulation, was observed in miR-DKO females. When the ovaries of these mice were examined, they showed a reduced number of ovulated oocytes (see panel b). But when miR-DKO mice were treated with hormones (gonadotropins) to stimulate ovulation, they produced a similar number of oocytes as heterozygous females.

Finally, a series of experiments were carried out to demonstrate that loss of miR-200b and miR-429 led to increased expression of the *Zeb1* gene in the

(continued)

Modern Approaches to Understanding Gene Function —continued



(a) Pregnancy rate of wild-type, miR-200b/miR-429 heterozygous females and of miR-DKO female mice crossed with miR-200b/miR-429 heterozygous male mice. (n indicates the number of female mice tested in each genotype; numbers above bars indicate pregnancies per successful coitus.) (b) Number of ovulated eggs derived from heterozygous and miR-DKO mice naturally or following induced ovulation. (*** $P < 0.001$)

pituitary gland, which directly resulted in decreased expression of the *Lhb* gene, which encodes the LH protein. *Zeb1* encodes a transcription factor thought to regulate expression of the *Lhb* gene.

Conclusions:

The ovulation experiment demonstrated that miR-DKO females could undergo normal oogenesis and ovulate when supplemented with hormones. This suggests that knocking out miR-200b and miR-429 impairs some aspect of the hormonal regulation of ovulation, most likely involving LH. Because the miR-200b cluster is present in humans and is adjacent to the human *ZEB1* and *LHB* genes, it is possible that miR-200b and miR-429 play a role in regulating ovulation in humans as well as mice.

Reference:

Hasuwa, H., Ueda, J., Ikawa, M., and Okabe, M. (2013). miR-200b and miR-429 Function in Mouse Ovulation and Are Essential for Female Fertility. *Science* 341:71–73.

Questions to Consider:

1. Based on the results described, draw a diagram to explain possible relationships between miR-220b and miR-429, the *Zeb1* and *Lhb* genes, and ovulation in miR-DKO female mice.
2. Another experiment in this study was to introduce a transgene that overexpresses miR-220b and miR-429 into the miR-DKO mice. Would you expect that the pregnancy rate of miR-DKO females expressing this transgene would increase, decrease, or stay the same as miR-DKO female mice without the transgene? Refer to the reference to examine the results produced by this mating.
3. Using OMIM (Online Mendelian Inheritance in Man) Web site, locate the human homolog for *Zeb1* and find out if there are any human disorders associated with problems with this gene. Do the results surprise you?

Mechanisms of RNA Interference

The ncRNAs involved in RNAi, broadly termed **small noncoding RNAs (sncRNAs),** are double-stranded RNAs that are 20–31 nucleotides long with 2-nucleotide overhangs at their 3′ ends. There are two sub-types of sncRNAs: **small interfering RNAs (siRNAs)** and the microRNAs (miRNAs) mentioned above. Although they arise from different sources, their mechanisms of action are similar.

Small Interfering RNAs siRNAs are derived from longer double-stranded RNA (dsRNA) molecules. These long dsRNAs may appear within cells as a result of virus infection or the expression of transposons, also called "jumping genes" (see Chapter 15), both of which may synthesize dsRNAs as part of their life cycles. RNAi may have evolved as a mechanism to recognize these dsRNAs and inactivate them, protecting the cell from external (viral) or internal (transposon) assaults. siRNAs can also be derived from lab-synthesized dsRNAs and introduced into cells for research or therapeutic purposes.

Whatever the source, when long dsRNAs are present in the cytoplasm of a eukaryotic cell, they are cleaved into approximately 22-nucleotide-long double-stranded siRNAs by an enzyme called **Dicer** (Figure 18.7, left). These siRNAs then associate with the **RNA-induced silencing complex (RISC).** RISC contains an **Argonaute** family protein that binds RNA and has endoribonuclease or "slicer" activity. RISC cleaves and evicts one of the two strands of the double-stranded siRNA and retains the other strand as a single-stranded siRNA "guide" to recruit RISC to a

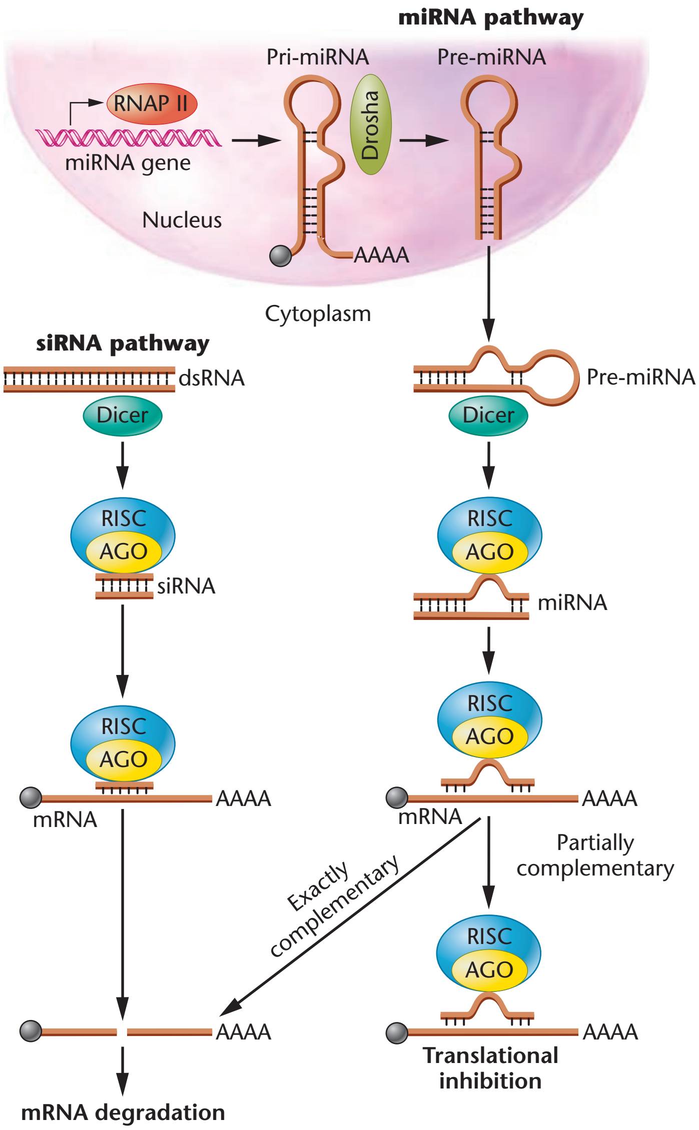


FIGURE 18.7 RNA interference pathways. Left: Double-stranded RNA is processed into siRNAs by Dicer. siRNAs then associate with RISC containing an Argonaute (AGO) family protein. RISC unwinds the siRNAs into single-stranded siRNAs and cleaves mRNAs complementary to the siRNA. Right: miRNA genes are transcribed as primary-miRNAs (pri-miRNAs), which are trimmed at the 5′and 3′ends by the nuclear enzyme Drosha to form pre-miRNAs, which are exported to the cytoplasm and processed by Dicer. These miRNAs then associate with RISC and mRNAs. If the miRNA and mRNA are perfectly complementary, the mRNA is destroyed; if there is a partial match, translation is inhibited.

complementary mRNA. RISC then cleaves the mRNA in the middle of the region of siRNA–mRNA complementarity (Figure 18.7, left, bottom). Cleaved mRNA fragments lacking a cap or a poly-A tail are then quickly degraded in the cell by exoribonucleases.

microRNAs miRNAs originate in the nucleus. They are transcribed from miRNA genes, which include self-complementary sequences. The initial transcripts, called **primary miRNAs (pri-miRNAs),** are processed similarly to mRNAs; they receive a cap and a poly-A tail, and some contain introns that are spliced out. However, because of their self-complementary sequences, pri-miRNAs form hairpin structures. A nuclear enzyme called **Drosha** removes the noncomplementary 5′ and 3′ ends to produce **pre-miRNAs** (Figure 18.7, right). These hairpins are then exported to the cytoplasm where they are cleaved by Dicer to produce mature double-stranded miRNAs and further processed to single-stranded miRNAs by RISC. Like siRNAs, miRNAs associate with RISC to target complementary sequences on mRNAs. Such complementary sequences serve as binding sites or **miRNA response elements (MREs).** MREs are commonly found in 3′ UTRs of mRNAs but can also be found in the 5′ UTRs or the coding region. If the miRNA–mRNA match is perfect (common in plants), the target mRNA is cleaved by RISC and degraded. But if the miRNA–mRNA match is partial (common in animals), it blocks translation (Figure 18.7, right, bottom).

miRNAs are found in plants and animals, and are encoded by some viruses. Studies suggest that there are at least 1500 miRNA genes in the human genome. However, one recent study analyzed RNA sequencing data from 13 different human tissue types and concluded that the total number of miRNAs is closer to 5000. miRNAs are not only plentiful but also have diverse roles. They have been shown to regulate genes involved in such processes as stress responses in plants, development in *C. elegans*, and cell-cycle control in mammalian cells.

Why is miRNA-mediated posttranscriptional regulation so widespread? Wouldn't it be more efficient for the cell to repress transcription rather than deploy an miRNA to destroy or inhibit the mRNA? The answer to these questions partly lies in the fact that mRNAs may be translated many times after transcription is stopped.

To achieve a rapid change in gene expression, a cell can turn off transcription and deploy an miRNA to target the existing mRNAs in the cytoplasm. For example, miRNAs are key regulators in mammalian embryonic stem cells (ES cells), the cells of the embryo that give rise to all the differentiated cell types of adult tissues. ES cells express "stemness" genes (such as *Sox2*, *Oct4*, *KLF4*, *Lin28*, and *Myc*), which suppress differentiation and promote stem cell maintenance. Loss of these genes results in differentiation of ES cells, while persistent activity results in an inability to produce specialized cells—both scenarios are lethal. To enable differentiation, ES daughter cells express miR-145 and let-7 miRNAs, which target and downregulate stemness mRNAs.

A better understanding of miRNA regulation of cellular differentiation is likely to improve stem cell therapy by ensuring that patients receive cells that properly differentiate into the desired cell types rather than fail to differentiate and pose a risk for tumor formation—a potential unwanted side effect of stem cell therapy.

NOW SOLVE THIS

18.2 Some scientists use the analogy that the RNA-induced silencing complex (RISC) is a "programmable search engine" that uses miRNAs as programs. What does RISC search for, and how does an miRNA "program" the search? What does RISC do when it finds what it is searching for?

- **HINT:** *The important concept here is that complementary base pairing enables an miRNA to guide RISC to its target for post-transcriptional regulation.*

RNA Interference in Research, Biotechnology, and Medicine

The discovery of the basic mechanism of RNAi in 1998 led to an almost immediate revolution in the investigation of gene function. As long as a gene's sequence is known, one can quickly synthesize dsRNAs corresponding to that sequence to inhibit or "knockdown" that gene's function. Just five years after the discovery of RNAi, Julie Ahringer's lab used RNAi on a massive scale to determine the loss-of-function phenotypes for 86 percent of the genes in the *C. elegans* genome! More recently, several tools have been developed for rapid and inexpensive RNAi-based gene investigation. Several biotechnology companies manufacture libraries of siRNA molecules for use in research that can be introduced into cultured cells to knock down specific gene products.

In addition to its use in research, RNAi is being developed as a potential pharmaceutical agent. In theory, any disease caused by overexpression of a specific gene, or even normal expression of an abnormal gene product, could be ameliorated by RNAi. Following some early difficulties with clinical trials, there has been a recent resurgence of RNAi-based drug development. This has been due, at least in part, to advances in siRNA delivery methods, such as nanoparticles and cell-penetrating peptides. In 2016, there were 20 ongoing clinical trials to treat viral infections like hepatitis and Ebola, as well as cancers, eye diseases, hemophilia, hypercholesterolemia, and even alcoholism. Some of these trials were in phase III, meaning that they were being administered to large groups to confirm effectiveness and monitor side effects.

One specific RNAi-based drug that has received a lot of attention lately is patisiran by Alnylam Pharmaceuticals. Patisiran is a siRNA and nanoparticle delivery system treats transthyretin amyloidosis, a disorder characterized by nervous system and cardiac problems due to a buildup of a mutant form of the transthyretin (TTR) protein. A phase II clinical trial showed that patisiran treatment results in a roughly 80 percent knockdown of TTR protein levels, suggesting effective targeting of the mRNA. Importantly, over 70 percent of patients either improved or stabilized their condition over the two-year study, with minimal adverse side effects. In September of 2017, Alnylam Pharmaceuticals announced positive results for a phase III clinical trial of patisiran. This is the first RNAi-based therapeutic to have a successful phase III clinical trial. Alnylam Pharmaceuticals expects that patisiran will receive approval from the U.S. Food and Drug Administration in 2018.

Long Noncoding RNAs and Posttranscriptional Regulation

In addition to sncRNAs discussed above, eukaryotic genomes also encode many **long noncoding RNAs (lncRNAs).** One obvious distinction is that lncRNAs are longer than sncRNAs and are often arbitrarily designated to be greater than 200 nucleotides in length. lncRNAs are produced in a similar fashion to mRNAs; they are modified with a cap and a poly-A tail, and they can be spliced. In contrast to mRNAs, they have no start and stop codons, indicating that they do not encode protein. The conservative estimate is that the human genome encodes ~17,000 lncRNAs.

lncRNAs have been linked to diverse regulatory functions. Some lncRNAs bind to chromatin-regulating complexes to influence chromatin modifications and alter patterns of gene expression (see Chapter 19). Others regulate transcription by directly associating with transcription factors. However, in this section we will focus on the *post-transcriptional* roles of lncRNAs.

When lncRNAs are complementary to mRNAs or pre-mRNAs, the two can hybridize by base pairing. In some cases, this leads to regulation of alternative splicing. For example, an lncRNA that binds to splice sites for an exon can lead to its exclusion from the mature transcript. In other cases, lncRNA–mRNA hybridization produces a dsRNA that triggers an RNAi response. It is processed by Dicer into siRNAs that then target complementary mRNAs for destruction by RISC (see Figure 18.7). Other studies show that lncRNAs can bind to mRNAs in ways that regulate their stability, decay, and translation.

Some lncRNAs function as **competing endogenous RNAs (ceRNAs).** Conceptually, ceRNAs are "sponges" that "soak up" miRNAs due to the presence of complementary miRNA-binding sites in their sequence—the miRNA response elements (MREs) introduced earlier in this chapter. Thus, ceRNAs compete with mRNAs for miRNA-binding. Whereas miRNAs downregulate their mRNA targets, ceRNAs are able to "derepress" the mRNA targets by sequestering miRNAs away from them. In other words, ceRNAs are decoys. The efficacy of a ceRNA depends on variables such as how many MREs it contains, how many copies of the ceRNA are expressed in the cell, and the affinity of its MREs for the miRNA.

Recent studies from Irene Bozzoni's group demonstrated that a ceRNA is important for the differentiation of muscle cells in mice and humans. Undifferentiated muscle cells, known as myoblasts, express an lncRNA called long intergenic noncoding RNA muscle differentiation 1 (linc-MD1) that contains 36 predicted MREs. Among them are MREs complementary to two specific miRNAs: miR-133 and miR-135. These miRNAs target mRNAs that encode two transcription factors involved in muscle differentiation—MEF2C and MAML1. This study showed that linc-MD1 acts as a ceRNA, binding and sequestering miR-133 and miR-135 away from their mRNA targets, leading to increased translation of the two transcription factors.

This study also showed that linc-MD1 expression levels are aberrantly low in the myoblasts of patients with Duchenne muscular dystrophy (DMD), a disease characterized by defects in muscle differentiation. Remarkably, when linc-MD1 was introduced into cultured myoblasts from DMD patients, muscle cell differentiation was partially restored. It appears that muscle cell differentiation is, in part, controlled by linc-MD1-mediated downregulation of the two miRNAs to allow the upregulation of the two muscle-specific transcription factors (**Figure 18.8**).

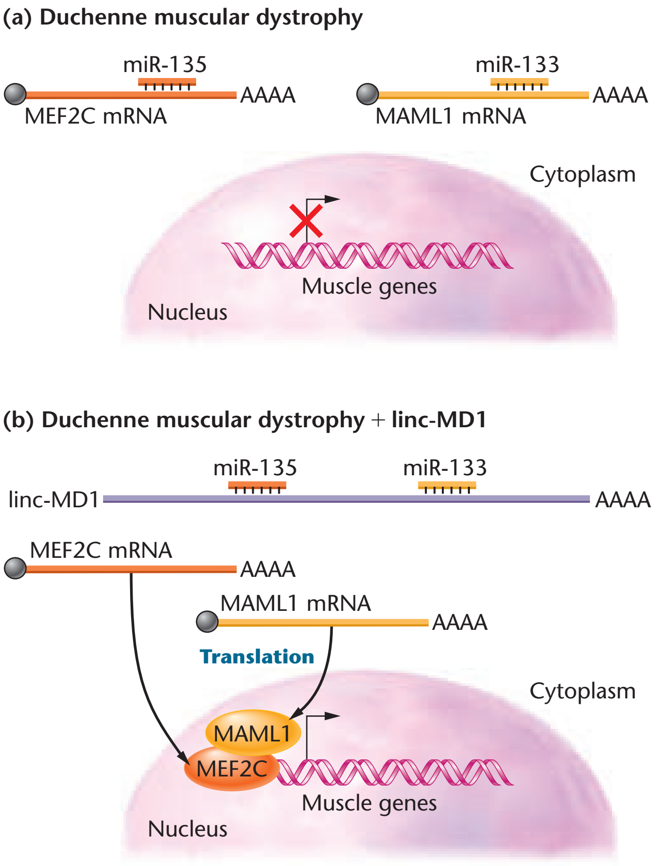


FIGURE 18.8 A lncRNA regulates muscle differentiation by sequestering miRNAs. (a) In Duchenne muscular dystrophy cells, linc-MD1 expression is aberrantly low and miR-133 and miR-135 downregulate mRNAs for the muscle-promoting transcription factors MEF2C and MAML1. (b) When linc-MD1 is introduced (using a virus vector), the miRNAs are bound and MEF2C and MAML1 mRNAs are upregulated, leading to increased muscle differentiation.

Circular RNAs

Another type of RNA that can compete for miRNA binding is the **circular RNA (circRNA).** These RNAs adopt a closed circular configuration with no free 5′-phosphate or 3′-OH ends, which makes them resistant to exoribonuclease attack. circRNAs were first observed in the cytoplasm of eukaryotic cells in 1979 using electron microscopy, but such studies offered no clues to their biological function. However, two recent studies have demonstrated that some circRNAs act as ceRNAs, or miRNA sponges.

One circRNA, called CDR1as/ciRS-7, is a sponge for miR-7 miRNAs. Expression of CDR1as/ciRS-7 leads to sequestration of miR-7 miRNAs and thus allows the expression of miR-7 target mRNAs. Given that the protein product of miR-7 targets α-synuclein mRNA, which encodes a protein that aggregates in patients with Parkinson disease, it is plausible that CDR1as/ciRS-7 plays an important role in the nervous system.

There are still many unanswered questions about circRNAs. How do they form? One hypothesis is that they originate during RNA splicing when the 3′ splice site for a downstream exon is spliced to the 5′ splice site of an upstream exon. Do circRNAs have other functions than acting as ceRNAs? This remains to be seen, but postulated functions include binding to sequester RNA-binding proteins (RBPs) from their mRNA targets, or serving as a "delivery vehicle" to shuttle RBPs to specific targets.

18.4 mRNA Localization and Translation Initiation Are Highly Regulated

We have already encountered several posttranscriptional mechanisms that impact translation. Alternative splicing determines which spliceforms may be translated, and mRNAs may be degraded or targeted by an miRNA to stop translation. However, translation may be regulated more directly as well. Even if an mRNA evades decay and is not targeted by an miRNA, it will not necessarily be translated by default. Moreover, translation may occur only in specific regions of the cell to restrict proteins to those areas. In this section we will look at how translation can be regulated to produce protein only when and where, within the cell, it is needed.

Cytoplasmic Polyadenylation

mRNAs are not always translated immediately; they may be stored for translation at a later time or in response to certain cues. For example, many mRNAs in an oocyte are held in a translationally repressed state until after fertilization. In this case, the cell is poised and ready for translation when the cell receives the signal to do so.

Many molecular mechanisms that regulate translation do so at the initiation step. Recall (Chapter 14) that translation initiation requires the formation of a closed-loop structure that involves eukaryotic initiation factors (eIFs) that assemble on the 5′ cap of the mRNA and interact with poly-A binding proteins (PABPs) on the poly-A tail. Thus, control of translation initiation often involves mechanisms that prevent the association of eIFs with one another or with the PABPs on the poly-A tail.

One well-studied mechanism used to control translation initiation involves regulation of the poly-A tail. mRNAs regulated in this manner often contain a *cis*-regulatory element in their 3′ UTR called a **cytoplasmic polyadenylation element (CPE).** The CPE sequence is often UUUUAU and is recognized and bound by an RNA-binding protein known as the **cytoplasmic polyadenylation element binding protein (CPEB)** [Figure 18.9(a)]. CPEB recruits **poly-A-specific ribonuclease (PARN)** to CPE-containing mRNAs, which shortens the poly-A tail to only ~40 adenosine residues. These shortened poly-A tails are bound by fewer molecules of PABP—too few to facilitate a stable interaction with the eIFs required for translation initiation. CPEB also recruits a protein called **Maskin** to CPE-containing mRNAs. Maskin binds to the cap binding protein, eIF4E, and blocks its interaction with eIF4G, which is necessary for translation initiation. Translationally repressed CPE-containing mRNAs are stored within cytoplasmic RNA-protein complexes known as **ribonucleoprotein (RNP) particles or granules,** which contain many RNA-binding proteins and translational regulators.

When a cell receives a signal to reactivate CPE-containing mRNAs, CPEB is phosphorylated by kinases [Figure 18.9(b)]. Phosphorylation induces a conformational change such that CPEB no longer binds to PARN and allows for a cytoplasmic poly-A polymerase to lengthen the poly-A tail. This longer poly-A tail is bound by additional PABPs, which displace Maskin and enable the formation of a stable translation initiation complex. Thus, CPE-mediated translation can be stimulated by kinase activity—a control mechanism that has been linked to a variety of cellular events such as cell-cycle control, hormone signaling, and neuron signaling.

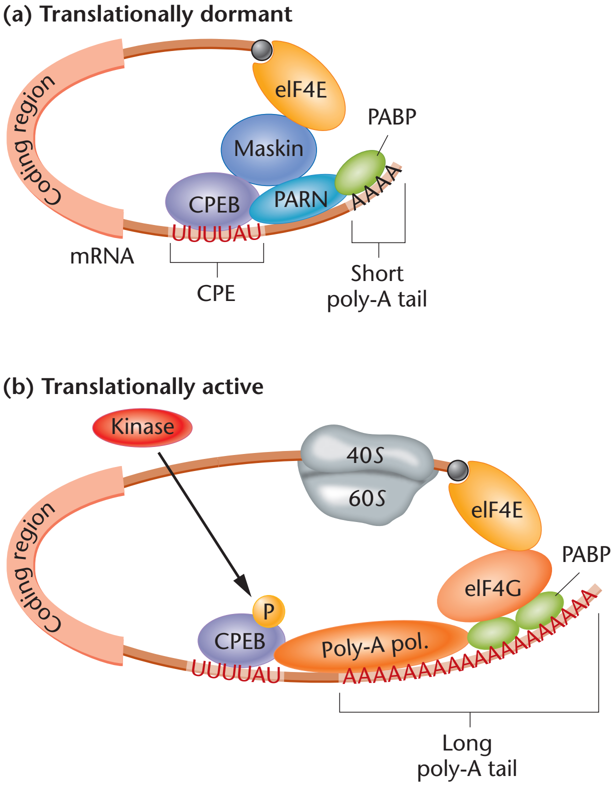


FIGURE 18.9 The control of translation by cytoplasmic polyadenylation.

mRNA Localization and Localized Translational Control

It has become clear that some mRNAs are localized to discrete regions within the cell and then are locally translated. These mechanisms make possible asymmetric protein distributions within the cell that define cellular regions with distinct functions. For example, specialized proteins localized in the highly branched dendrites of a neuron enable them to receive sensory information, while the proteins present in the axon of a neuron mediate the release of neurotransmitters that signal to other cells.

Similar to other posttranscriptional mechanisms, regulation of mRNA localization and localized translational control are governed by *cis*-regulatory sequences on the mRNA and *trans*-acting RNA-binding proteins. We have already seen that RBPs regulate mRNA splicing, stability, and decay; RBPs also regulate mRNA localization and localized translation. One of the best-described RBP–mRNA interactions governs the localization and translational control of actin mRNAs in crawling cells.

Following injury, fibroblasts migrate to the site of the wound and assist in wound healing. Fibroblasts and many other types of migrating cells control their direction of movement by controlling where within the cell they polymerize new cytoskeletal actin microfilaments. The "leading edge" of the cell where this actin polymerization occurs is called the lamellipodium.

Elegant studies by Robert Singer's lab showed that actin mRNA is localized to lamellipodia, and that localization is dependent on a 54-nucleotide element in the actin mRNA 3′ UTR termed a **zip code.** The zip code is a *cis*-regulatory element that serves as a binding site for a RBP called **zip code binding protein 1 (ZBP1).** ZBP1 initially binds the zip code sequence element of actin mRNAs in the nucleus. Once exported to the cytoplasm, ZBP1 blocks translation initiation by preventing the association of the large subunit (60*S*) of the ribosome. In addition, ZBP1 associates with cytoskeleton motor proteins to facilitate the transport of the mRNA to the lamellipodium. Once the mRNAs arrive at the final destination, a kinase called **Src** phosphorylates ZBP1, which disrupts RNA binding and allows translation initiation (**Figure 18.10**).

Since Src activity is limited to the cell periphery, this mechanism allows the transport of actin mRNAs in a translationally repressed state to the cell periphery, thus controlling where actin will be translated and polymerize.

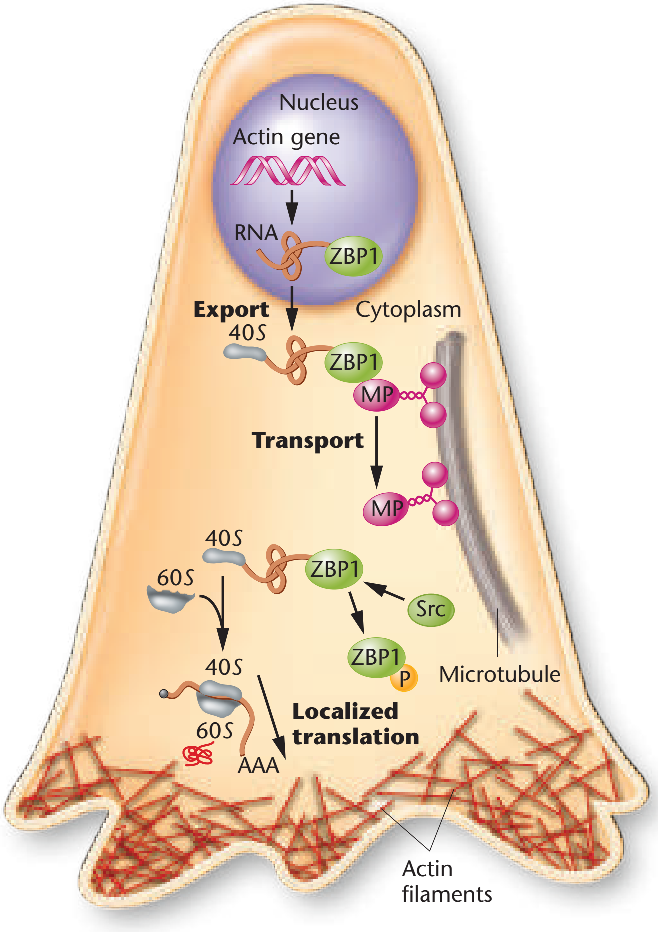


FIGURE 18.10 Localization and translational regulation of actin mRNA. The RNA-binding protein ZBP1 associates with actin mRNA in the nucleus and escorts it to the cytoplasm. ZBP1 blocks translation and binds cytoskeleton motor proteins (MP), which transport ZBP1 and actin mRNA to the cell periphery. At the cell periphery, ZBP1 is phosphorylated by Src and dissociates from actin mRNA, allowing it to be translated by a ribosome (40*S*/60*S*). Actin translation and polymerization at the leading edge direct cell movement.

Consistent with this model, mouse fibroblasts lacking ZBP1 have reduced actin mRNA localization and reduced directional motility.

In addition to crawling cells, there is some evidence suggesting that an actin mRNA localization mechanism is used in other cell types as well. For example, local translation of actin in the axons and dendrites of neurons is required for their guided outgrowth. Mouse neurons lacking ZBP1 have reduced dendrite length and exhibit defects in axon guidance. mRNA localization and translational control are particularly important for nervous system function. In fact, defects in mRNA localization in neurons have been implicated in human disorders such as fragile-X syndrome, spinal muscular atrophy, and spinocerebellar ataxia.

NOW SOLVE THIS

18.3 Consider the example that actin mRNA localization is important for fibroblast migration. What would you predict to be the consequence of deleting the zip code sequence element of the actin mRNA?

■ **HINT:** *The key to answering this question is recalling that the zip code is a* cis-*acting element that is bound by an RNA-binding protein involved in its localization and translational control.*

18.5 Posttranslational Modifications Regulate Protein Activity

Even after translation is complete, the activity of the gene products can still be regulated through a suite of **posttranslational modifications.** You've learned about several of these mechanisms earlier in the text (Chapter 14), such as the cleaving of the N-terminal amino acid (common for many eukaryotic proteins) and the association of proteins with prosthetic groups (such as the heme groups in the oxygen-carrying protein hemoglobin). In addition, proteins may be posttranslationally modified by the covalent attachment of various molecules. Such additions can change a protein's stability; subcellular localization; or affinity for other proteins, nucleic acids, or molecules. Since covalent modification is an enzyme-catalyzed event, the regulation of such enzymes is a critical step for controlling gene expression at the posttranslational level.

Regulation of Proteins by Phosphorylation

We do not know the full extent of posttranslational modifications within the proteome for any given species or cell type. However, a 2011 annotation of experimentally determined data in the SWISS-PROT curated protein sequence database has given us some insights into the frequency of many types of posttranslational modifications.

Phosphorylation is the most common type, accounting for approximately 65 percent of all analyzed posttranslational modifications. Phosphorylation is mediated by a class of enzymes called **kinases.** Kinases catalyze the addition of a phosphate group to serine, tyrosine, or threonine amino acid side chains. Such additions are reversible. **Phosphatases** are enzymes that remove phosphates. It is calculated that the human genome contains 518 kinase-encoding genes and 147 phosphatase-encoding genes. This suite of enzymes can be used in countless ways to regulate protein activity.

Phosphorylation usually induces conformational changes. These changes can have different effects depending on the type of target. For example, enzymes may be turned on or off by phosphorylation where conformational changes alter substrate binding. Transcription factors may be turned on or off by phosphorylation based on how the conformational changes impact its affinity for the target DNA sequence. In some cases, there is more than one phosphorylation site on a protein; phosphorylation in one site may activate, while phosphorylation at the other site may inactivate the protein.

Ubiquitin-Mediated Protein Degradation

One important way to regulate protein activity after translation is through the targeting of specific proteins for degradation. The principal mechanism by which the eukaryotic cell targets a protein for degradation is by covalently modifying it with **ubiquitin,** a small protein with 76 amino acids that is found in all eukaryotic cells. The fact that it is *ubiquitous* gives ubiquitin its name.

Ubiquitin is covalently attached to a target protein via a lysine side chain through a process called **ubiquitination.** Subsequently, lysine side chains in the attached ubiquitin molecule can be modified by the addition of other ubiquitin molecules. This process can be repeated to form long poly-ubiquitin chains, which serve as "tags" that mark the protein for destruction. Poly-ubiquitinated proteins are recognized by the **proteasome,** a multi-subunit protein complex with protease (protein cleaving) activity. The proteasome unwinds target proteins, removes their ubiquitin tags, and breaks the protein into small peptides about 7–8 amino acids long (**Figure 18.11**).

Since ubiquitinated proteins are quickly destroyed, the determination of which proteins get ubiquitinated is a major regulatory step. A class of enzymes, **ubiquitin ligases,** recognize and bind specific target proteins, and catalyze the processive addition of ubiquitin residues (Figure 18.11). In turn, ubiquitin ligase activity can be regulated in many ways to serve the needs of the cell. For the discovery of ubiquitin-mediated protein degradation, Aaron Ciechanover, Avram Hershko, and Irwin Rose were awarded the 2004 Nobel Prize in Chemistry.

An important example of ubiquitin-mediated protein degradation is that of the transcription factor **p53.** Sometimes called "the guardian of the genome," p53 plays an important role in protecting the cell against harm from DNA damage. If DNA damage occurs, p53 activates the transcription of genes that encode proteins that arrest the cell cycle and promote DNA repair. However, in healthy cells the level of p53 is kept low by degradation. A ubiquitin ligase called Mdm2 binds p53 and marks it for degradation by the proteasome. However, when the cell senses DNA damage, Mdm2 is phosphorylated, which causes it to lose its affinity for p53. Thus, p53 destruction is halted and its levels increase in the cell, allowing it to enact cell-cycle arrest and activation of DNA repair.

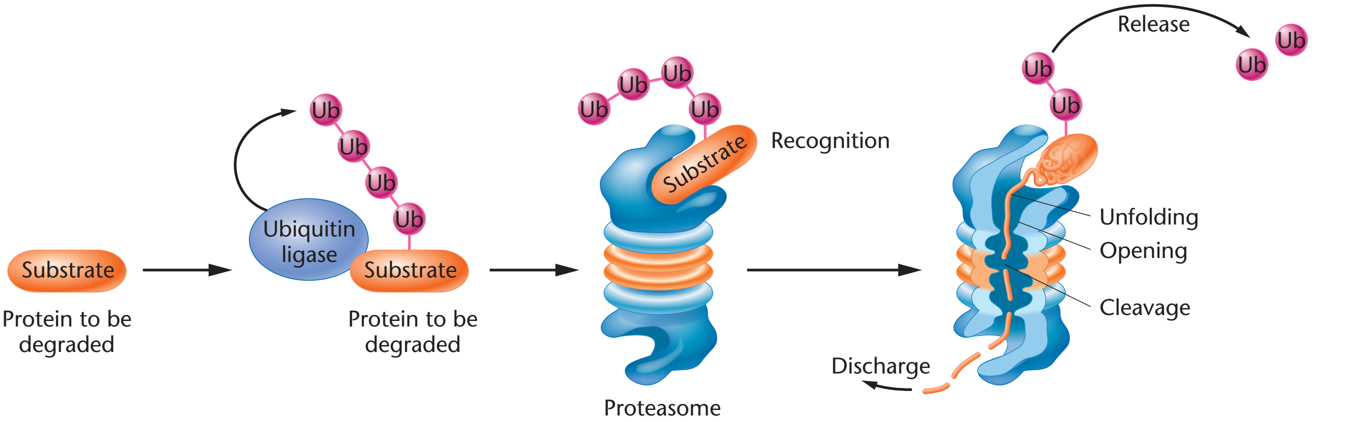


FIGURE 18.11 Ubiquitin-mediated protein degradation. Ubiquitin ligase enzymes recognize substrate proteins and catalyze the addition of ubiquitin (Ub) residues to create a long chain. Ubiquitinated proteins are then recognized by the proteasome, which removes ubiquitin tags, unfolds the protein, and proteolytically cleaves it into small polypeptides.

It is estimated that there are over 600 ubiquitin ligase—encoding genes in the human genome. Experimental determination of substrate proteins for any given ubiquitin ligase has proven difficult. However, a handful of recent studies suggest that some human ubiquitin ligases interact with over 40 different substrate proteins. Overall, scientists estimate that human ubiquitin ligases target over 9000 different proteins, which accounts for approximately 40 percent of the protein-coding genes in the human genome. This suggests that ubiquitin-mediated protein degradation may be a broadly used mechanism to regulate biological function.

GENETICS, ETHICS, AND SOCIETY

Is DNA Enough?

Genetic research has given us a wealth of information about the chemical basis of life. We know that DNA serves as the genetic material in all cellular life. In addition, we have the complete DNA sequences from many species. We understand how DNA is replicated, how it is expressed, how genes are regulated transcriptionally and post transcriptionally, how DNA is modified epigenetically, and how many of the RNAs and proteins encoded by DNA sequences act in numerous cellular processes and pathways.

But, can this vast knowledge about the molecular components of a living cell really tell us about how a cell or an organism works in the real world?

There are two sides to the debate about *what makes life work*. One side expresses confidence that DNA is enough. They argue that if you have the essential set of genes that encode all of the cellular components of an organism, those cells will naturally emerge from the expression of those genes. Proponents of the DNA-is-enough position point to the field of synthetic biology, whose ultimate goal is to create a living cell from its nonliving elements—perhaps as simple as a single DNA or RNA molecule enclosed in a membrane. In addition, researchers have produced computer simulations of a simple organism (*Mycoplasma genitalium*) by incorporating all known information about its DNA and proteins.

Research from the J. Craig Venter laboratory[1] has also excited many proponents of the DNA-is-enough position. As described later in the text (Chapter 22), the researchers synthesized the entire genome of a single-celled microbe *in vitro*, and then introduced the synthetic DNA into a different microbe that had its genome removed. The resulting cell reproduced and expressed many of the characteristics of its DNA parent. The study led to headlines that scientists had synthesized "artificial life" and the first "synthetic cell."

The other side in the debate argues that DNA sequences alone cannot define the complexities within a living cell, much less an entire multicellular organism. To explain a living entity, they say that one would need to understand the three-dimensional structures of all gene products and how these structures interact with each other, how cells converse with their environment, and how evolution works on each cell of an organism. They claim that cells and organisms are dynamic, constantly changing as they undergo development and interact with their surroundings. What constitutes a living cell in one set of circumstances will not work in another. Even attempts to define small gene networks for simple organisms (such as those described in Chapter 21) are fraught with immense complexity and hampered by incomplete knowledge of the components.

The DNA-is-insufficient side also points to Venter's research as support for their position. They explain that the synthetic cells created in Venter's lab are not new life forms created from the introduced DNA, but simply preexisting cells with prosthetic genomes. The synthesized DNA cannot function on its own outside a recipient cell, which contains all the components necessary to support the DNA sequence—such as the gene expression and regulation machinery, a metabolic apparatus, and a cellular structure. The DNA-is-insufficient side claims that trying to understand life by knowing DNA sequences is like trying to learn a language by memorizing a dictionary.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links that deal with the ethical and scientific arguments in the "Is DNA Enough?" debate.

1. What are some additional arguments that are put forward on both sides of this debate? After reading about this topic, decide which side you would support and explain why.

   To start your reading, go to Yong, E. (2012). Will We Ever . . . Reveal All the Secrets of Life from DNA? at http://www.bbc.com/future/story/20121102-will-we-ever-crack-lifes-code. *Also, see* Noireaux, V. et al. (2011). Development of an Artificial Cell, from Self-Organization to Computation and Self-Reproduction. *Proc. Natl. Acad. Sci.* 108:3473–3480.

2. The creation of synthetic cells triggers a number of ethical concerns. What are some of these considerations?

   Three major concerns are discussed in Douglas, T., and Savulescu, J. (2010). Synthetic Biology and the Ethics of Knowledge. *J. Med. Ethics* 36:687–693.

[1] Hutchison, C. A. III, et al. (2016). Design and Synthesis of a Minimal Bacterial Genome. *Science* 351:1414.

CASE STUDY A mysterious muscular dystrophy

A man in his early 30s suddenly developed weakness in his hands and neck, followed weeks later by burning muscle pain—all symptoms of late-onset muscular dystrophy. His internist ordered genetic tests to determine whether he had one of the most common adult-onset muscular dystrophies—myotonic dystrophy type 1 (DM1) or myotonic dystrophy type 2 (DM2). The tests detect mutations in the *DMPK* and *CNBP* genes, the only genes known to be associated with DM1 and DM2. While awaiting the results of the gene tests, the internist explained that the disease-causing mutations in these genes do not result in changes to the coding sequence. Rather, myotonic dystrophies result from increased, or expanded, numbers of tri- and tetranucleotide repeats in the 3′ untranslated region of the *DMPK* or *CNBP* genes. The doctor went on to explain that the presence of RNAs with expanded numbers of repeats leads to aberrant alternative splicing of other mRNAs, causing widespread disruption of cellular pathways. This discussion raises a number of interesting questions.

1. What is alternative splicing, where does it occur, and how could disrupting it affect the expression of the affected gene(s)?
2. What role might the expanded tri- and tetranucleotide repeats play in the altered splicing?
3. DM1 is characterized by a phenomenon known as genetic anticipation (see Chapter 4) where the age of onset tends to decrease and the severity of the symptoms tend to increase from one generation to the next due to expansion of the trinucleotide repeats in the *DMPK* gene. What are the implications of a diagnosis of DM1 in this patient with respect to his 4-year-old son, and 2-year-old daughter?

For related reading, see Pavićević, D. S., et al. (2013). Molecular Genetics and Genetic Testing in Myotonic Dystrophy Type 1. *Biomed. Res. Int.* 2013, 391821.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. In eukaryotes, posttranscriptional gene regulation can occur at any of the steps from nuclear RNA processing to posttranslational modification of proteins.
2. The pre-mRNAs of many eukaryotic genes undergo alternative splicing to produce different spliceforms encoding different protein isoforms, which may have different functions.
3. Defects in alternative splicing are associated with several human diseases.
4. Modulation of mRNA stability and decay can eliminate aberrant mRNAs and regulate gene expression.
5. Noncoding RNAs, such as microRNAs (miRNAs), can mediate sequence-specific degradation or translational inhibition of target mRNAs in a process called RNA interference (RNAi).
6. RNAi has been harnessed as an important tool for research and biotechnology and is currently in clinical trials for medical applications.
7. Some mRNAs are stored in an inactive state for translation at a later time or in response to specific cues.
8. mRNAs may be localized to specific regions of the cell and then locally translated to create asymmetric protein distribution within the eukaryotic cell.
9. Following translation, protein activity can be modulated by posttranslational modifications, such as phosphorylation or ubiquitin-mediated degradation.

INSIGHTS AND SOLUTIONS

Scientists estimate that more than 15 percent of disease-causing mutations involve errors in alternative splicing. However, there is an interesting case in which a mutation that deletes an exon results in increased protein production. Mutations that delete exon 45 of the 79-exon *dystrophin* gene are the most common cause of Duchenne muscular dystrophy (DMD), a disease associated with progressive muscle degeneration. However, some individuals with deletions of both exons 45 and 46 have Becker muscular dystrophy (BMD), a milder form of muscular dystrophy. Provide a possible explanation for why BMD patients, with a deletion of both exon 45 and 46, produce more dystrophin than DMD patients do.

Solution: Having a deletion of one exon has several possible effects on a gene product. One possibility is that the mRNA transcribed from the exon-deleted *dystrophin* gene is unstable, leading to a lack of dystrophin protein production. Even if the mRNA is stable, the resulting mutated dystrophin protein could be targeted for rapid degradation, leading to the absence of stable active protein. Another possibility is that the deletion of one exon creates a frameshift leading to a premature stop codon. As the *dystrophin* gene has 79 exons spanning over 2.6 million base pairs, a frameshift in exon 46 could create a stop codon near the middle of the gene, which would trigger nonsense-mediated mRNA decay. Any mRNA escaping degradation would encode a shorter than normal dystrophin protein, which would likely be nonfunctional.

It is possible that a deletion encompassing both exon 45 and 46 could restore the reading frame of the dystrophin protein in exon 47. The protein product of this gene would be missing amino acid sequences encoded by the two missing exons; however, the protein itself could still have some activity, partially preserving the wild-type phenotype.

This concept—skipping an exon to restore a frameshift to the proper reading frame—is the focus of a medical treatment for DMD [van Deutekom and van Ommen (2003). *Nat. Rev. Genetics* 4:774–783]. See also the earlier Genetics, Ethics, and Society essay (Chapter 13) that discusses a recently approved drug that ameliorates symptoms in DMD patients with mutations in a specific region of the *dystrophin* gene.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on how eukaryotic gene expression is regulated posttranscriptionally. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter:
(a) How do we know that alternative splicing enables one gene to encode different isoforms with different functions?
(b) How do we know that misregulation of mRNA stability and decay is a contributing factor in some cancers?
(c) How do we know that double-stranded RNA molecules can control gene expression?
(d) How do we know that microRNAs negatively regulate target mRNAs?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page XXX. The third concept describes how the interaction between *cis*-acting sequence elements on mRNA and *trans*-acting RNA-binding proteins (RBPs) regulates mRNAs and gene expression. Write a short essay describing how an mRNA may be regulated in three different ways by specific *cis*-elements and RBPs.

3. List three types of alternative splicing patterns and how they lead to the production of different protein isoforms.

4. Consider the *CT/CGRP* example of alternative splicing shown in Figure 18.3. Which different types of alternative splicing patterns are represented?

5. Explain how the use of alternative promoters and alternative polyadenylation signals produces mRNAs with different 5′ and 3′ ends.

6. Explain how a tissue-specific RNA-binding protein can lead to tissue-specific alternative splicing via splicing enhancers or splicing silencers.

7. The regulation of mRNA decay relies heavily upon deadenylases and decapping enzymes. Explain how these classes of enzymes are critical to initiating mRNA decay.

8. Nonsense-mediated decay is an mRNA surveillance pathway that eliminates mRNAs with premature stop codons. How does the cell distinguish between normal mRNAs and those with a premature stop?

9. AU-rich elements (AREs) are *cis*-elements in mRNAs that regulate stability and decay. How is it possible that a single mRNA sequence element can serve to stabilize an mRNA in some cases and lead to its decay in other scenarios?

10. What are processing bodies (P bodies), and what role do they play in mRNA regulation?

11. In 1998, future Nobel laureates Andrew Fire and Craig Mello, and colleagues, published an article in *Nature* entitled, “Potent and Specific Genetic Interference by Double-Stranded RNA in *Caenorhabditis elegans*.” Explain how RNAi is both “potent and specific.”

12. Present an overview of RNA interference (RNAi). How does the silencing process begin, and what major components participate?

13. RNAi may be directed by small interfering RNAs (siRNAs) or microRNAs (miRNAs); how are these similar, and how are they different?

14. miRNAs target endogenous mRNAs in a sequence-specific manner. Explain, conceptually, how one might identify potential mRNA targets for a given miRNA if you only know the sequence of the miRNA and the sequence of all mRNAs in a cell or tissue of interest.

15. In principle, RNAi may be used to fight viral infection. How might this work?

16. Competing endogenous RNAs act as molecular “sponges.” What does this mean, and what do they compete with?

17. While circular RNAs were first described long ago, they have only recently been investigated for function. What are their known and suspected functions in the cell?

18. How are mRNAs stored within the cell in a translationally inactive state, and how can their translation be stimulated?

19. How and why are eukaryotic mRNAs transported and localized to discrete regions of the cell?

20. How is it possible that a given mRNA in a cell is found throughout the cytoplasm but the protein that it encodes is only found in a few specific regions?

21. How may the covalent modification of a protein with a phosphate group alter its function?

22. What role do ubiquitin ligases play in the regulation of gene expression?

23. The proteasome is a multi-subunit machine that unfolds and degrades proteins. How is its activity regulated such that it only degrades certain proteins?

Extra-Spicy Problems

24. In this chapter, we discussed several specific *cis*-elements in mRNAs that regulate splicing, stability, decay, localization, and translation. However, it is likely that many other uncharacterized *cis*-elements exist. One way in which they may be characterized is through the use of a reporter gene such as the gene encoding the green fluorescent protein (GFP) from jellyfish. GFP emits green fluorescence when excited by blue light. Explain how one might be able to devise an assay to test for the effect of various *cis*-elements on posttranscriptional gene regulation using cells that transcribe a GFP mRNA with genetically inserted *cis*-elements.

25. Incorrectly spliced RNAs often lead to human pathologies. Scientists have examined cancer cells for splice-specific changes and found that many of the changes disrupt tumor-suppressor gene function [Xu and Lee (2003). *Nucl. Acids Res.* 31:5635–5643]. In general, what would be the effects of splicing changes on these RNAs and the function of tumor-suppressor gene function? How might loss of splicing specificity be associated with cancer?

26. Mutations in the *low-density lipoprotein receptor* (*LDLR*) gene are a primary cause of familial hypercholesterolemia. One such mutation is a SNP in exon 12 of the *LDLR*. In premenopausal women, but not in men or postmenopausal women, this SNP leads to skipping of exon 12 and production of a truncated nonfunctional protein. It is hypothesized that this SNP compromises a splice enhancer [Zhu et al. (2007). *Hum Mol Genet.* 16:1765–1772]. What are some possible ways in which this SNP can lead to this defect, but only in premenopausal women?

27. RNA helicases are a class of proteins that bind mRNAs and influence their secondary structures and interactions with other proteins. RNA helicases have been implicated in many steps of RNA regulation such as splicing, decay, and translation. Why might these enzymes be so ubiquitously required for RNA regulation?

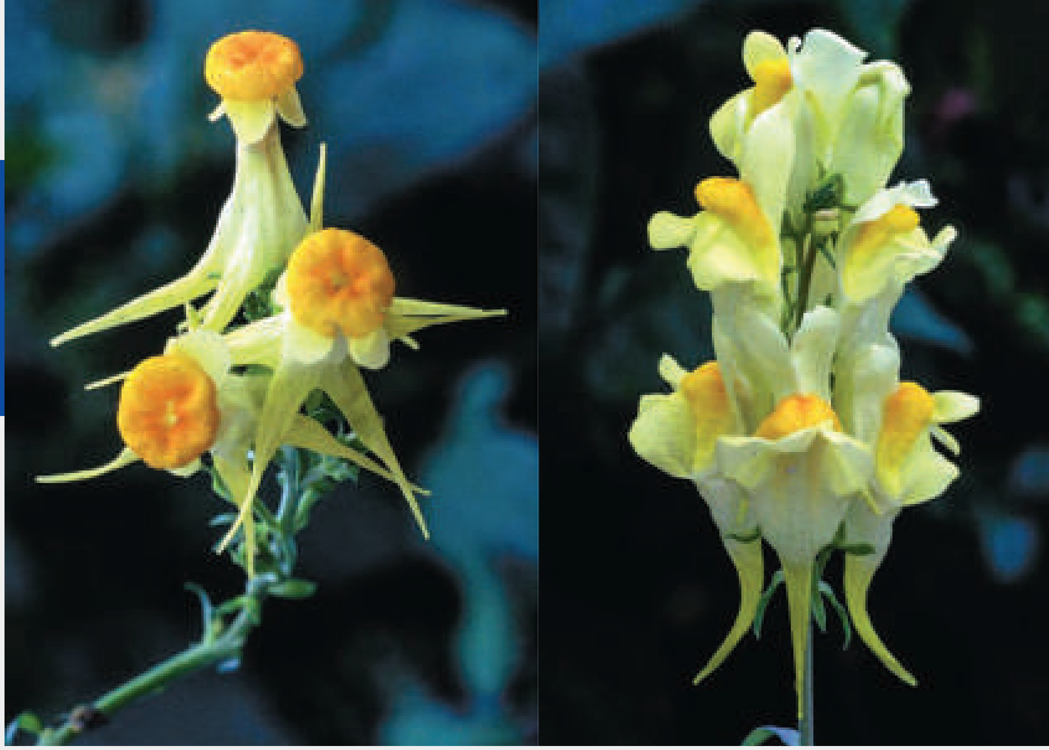
28. While miRNA response elements (MREs) may be located anywhere within an mRNA, they are most often found outside the coding region in the 5′ or 3′ UTR. Explain why this is likely the case given that miRNAs often target more than one mRNA.

29. RNAi is currently being tested as a therapeutic tool for genetic diseases and other conditions. Consider the following: cystic fibrosis caused by loss of function of the *CFTR* gene, HIV infection, and cancer caused by hyperactivity of a growth factor receptor. Which of these may be treatable by RNAi, and which not? Explain your reasoning.

30. The localization and translational control of actin mRNA is important for the migration of fibroblasts and is regulated by the activity of the kinase Src (see Figure 18.10). Src is activated by phosphorylation when cell surface receptors bind to signaling molecules. How might this system lead to a cell migrating in a specific direction? How might the cell migrate away from repulsive signals?

31. Explain how the expression of a single gene can be quickly, efficiently, and specifically shut down at the transcriptional, posttranscriptional, and posttranslational stages through the coordinated expression of a transcriptional repressor, an miRNA, and a ubiquitin ligase.

19



Epigenetic Regulation of Gene Expression

In toadflax, the shape of individual flowers changes from bilateral symmetry (photo on the left) to radial symmetry (photo on the right) in a naturally occurring, heritable gene silencing epimutation associated with the methylation of a single gene. There is no alteration of the DNA sequence at this locus.

CHAPTER CONCEPTS

- Genomes may be altered without changing the DNA sequence by several epigenetic mechanisms that include DNA methylation and demethylation, histone modifications, and the action of noncoding RNA molecules.
- The epigenome represents a specific pattern of epigenetic modifications present in a cell at a given time. Over its lifetime, a cell will have only one genome but will exist in many epigenomic states.
- Combinations of covalent histone modifications control the transcriptional status of chromatin regions, either activating or silencing genes. The patterns of histone modification that regulate gene expression are referred to as the histone code.
- Epigenetic mechanisms can limit transcriptional activity to only one parental allele in a parent-specific pattern called imprinting, or act at random in a wide array of autosomal genes, called monoallelic epigenetic expression.
- Abnormalities of DNA methylation patterns are a hallmark of cancer. Hypomethylation (undermethylation) activates many genes that are normally inactive, including oncogenes, which trigger uncontrolled cell division. Hypermethylation (overmethylation) of other genome regions facilitates chromatin remodeling, chromosome rearrangements, and changes in chromosome number. In addition to changes in DNA methylation, cancers also exhibit disruptions in the patterns of covalent histone modifications.
- Epigenetic changes to the genome are reversible, and therefore, the enzymes involved in adding or removing chemical groups on DNA and histones are the focus of new methods of chemotherapy.
- The epigenomic state of cells mediates the interaction between the external environment and the genome. These environmental cues can be physical or behavioral, and some of the new patterns of gene expression result in changes in heritable traits.

In previous chapters we established that gene expression in eukaryotes can be regulated both transcriptionally (Chapter 17) and posttranscriptionally (Chapter 18). However, as we have learned more about genome organization and the regulation of gene expression, it is clear that classical regulatory mechanisms cannot fully explain how some phenotypes arise. For example, monozygotic twins have identical genotypes but often have different phenotypes. In other instances, although one allele of each gene is inherited maternally and one is inherited paternally, only the maternal or paternal allele is expressed, while the other remains transcriptionally silent. Investigations of such phenomena have led to the emerging field of epigenetics, which is providing us with a molecular basis for understanding how heritable genomic alterations other than those encoded in the DNA sequence can influence phenotypic variation (**Figure 19.1**).

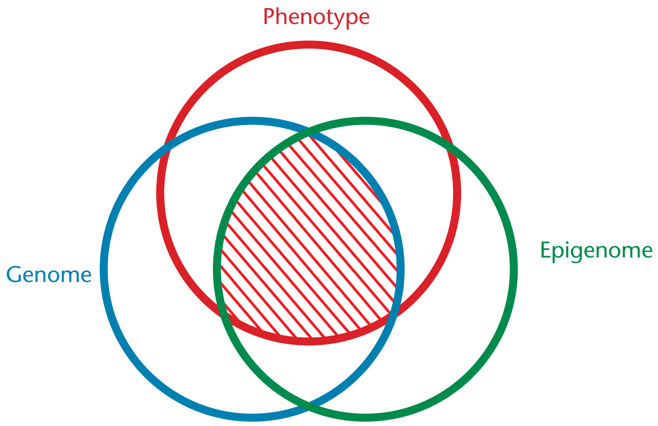


FIGURE 19.1 The phenotype of an organism is the product of interactions between the genome and the epigenome (hatched areas). The genome is constant from fertilization throughout life, but cells, tissues, and the organism develop different epigenomes as a result of epigenetic reprogramming of gene activity in response to environmental stimuli. These reprogramming events lead to phenotypic changes throughout the life cycle.

Epigenetics can be defined as the study of phenomena and mechanisms that cause chromosome-associated heritable changes to gene expression that are not dependent on changes in DNA sequence. The **epigenome** refers to the specific pattern of epigenetic modifications present in a cell at a given period of time. During its life span, an organism has one genome, which can be modified at different times to produce many different epigenomic states.

Knowledge of the mechanisms of epigenetic modifications to the genome, how these modifications are maintained and transmitted, and their relationship to basic biological processes is important to enhance our understanding of reproduction and development, disease processes, and the evolution of adaptations to the environment, including behavior.

Current research efforts are focused on several aspects of epigenetics: (1) how an epigenomic state arises in developing and differentiated cells and (2) how these epigenetic states are transmitted via mitosis and meiosis, making them heritable traits. The fruits of these efforts will be a major focus of this chapter. In addition, because epigenetically controlled alterations to the genome are associated with common diseases such as cancer and diabetes, efforts are also directed toward developing drugs that can modify or reverse disease-associated epigenetic changes in cells.

19.1 Molecular Alterations to the Genome Create an Epigenome

Unlike the genome, which is identical in all cell types of an organism, the epigenome is cell-type specific and changes throughout the life cycle in response to environmental cues. Like the genome, the epigenome can be transmitted to daughter cells by mitosis and to future generations by meiosis. In this section, we will examine mechanisms that shape the epigenome.

There are three major epigenetic mechanisms: (1) reversible modification of DNA by the addition or removal of methyl groups; (2) chromatin remodeling by the addition or removal of chemical groups to histone proteins; and (3) regulation of gene expression by noncoding RNA molecules. We now will look at each of these molecular activities in turn.

DNA Methylation and the Methylome

The set of methylated nucleotides present in an organism's genome at a given time is known as the **methylome**. The methylome is cell and tissue specific, but is not fixed, and changes as cells are called upon to respond to changing conditions. In mammals, **DNA methylation** takes place after DNA replication and during cell differentiation. This process involves the addition of a methyl group ($-CH_3$) to cytosine on the 5-carbon of the cytosine nitrogenous base (Figure 19.2), resulting in 5-methylcytosine (5mC), a reaction catalyzed by a family of enzymes called DNA methyltransferases (DNMTs). In humans, 5mC is present in about 1.5 percent of the genomic DNA.

Methylation takes place almost exclusively on cytosine bases located adjacent to a guanine base, a combination called a CpG dinucleotide:

$$\begin{array}{l} 5' - \overset{m}{C}pG - 3' \\ 3' - Gp\underset{m}{C} - 5' \end{array}$$

Many of these dinucleotide sites are clustered in regions called CpG islands, located in promoter and upstream sequences [Figure 19.3(a)].

CpG islands and promoters adjacent to essential genes (housekeeping genes) and cell-specific genes are unmethylated, making them available for transcription. Genes with adjacent methylated CpG islands and methylated CpG sequences within promoters are transcriptionally silenced.

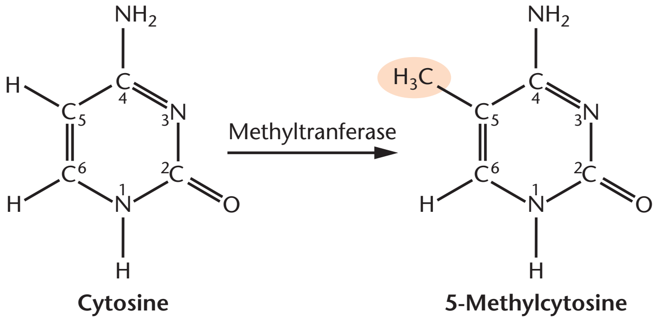


FIGURE 19.2 In DNA methylation, methyltransferase enzymes catalyze the transfer of a methyl group from a methyl donor to cytosine, producing 5-methylcytosine.

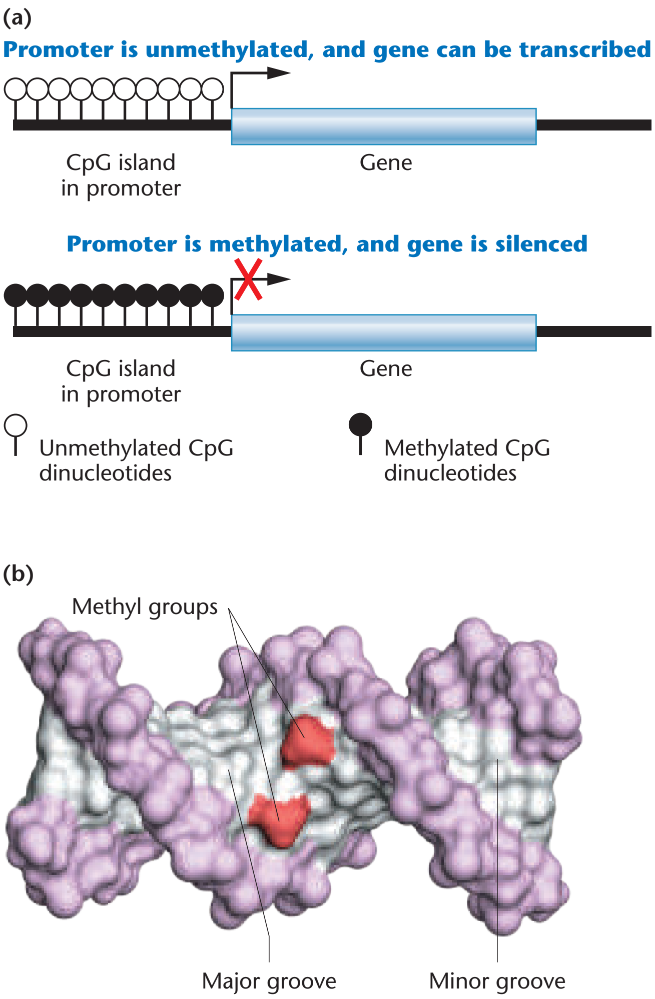

(a)
Promoter is unmethylated, and gene can be transcribed
CpG island in promoter
Gene
Promoter is methylated, and gene is silenced
CpG island in promoter
Gene
Unmethylated CpG dinucleotides
Methylated CpG dinucleotides
(b)
Methyl groups
Major groove
Minor groove

FIGURE 19.3 (a) Methylation patterns of CpG dinucleotides in promoters control activity of adjacent genes. (b) Methyl groups (highlighted in red) occupy the major groove of DNA and prevent binding of transcription factors, silencing genes.

The added methyl groups occupy the major groove of DNA and silence genes by blocking the binding of transcription factors and other proteins necessary to form transcription complexes [Figure 19.3(b)].

However, the bulk of methylated CpG dinucleotides are not adjacent to genes; instead they are located in the repetitive DNA sequences of heterochromatic regions of the genome, including the centromere. Methylation of these sequences contributes to silencing of transcription and replication of transposable elements such as LINE and SINE sequences which constitute a major portion of the human genome. (See Chapter 12 for a detailed discussion of repetitive sequences, heterochromatin, and chromosome organization.) Heterochromatic methylation is important in maintaining chromosome stability by preventing translocations and related abnormalities.

Three highly studied DNA methyltransferases (DNMTs) are involved in creating and maintaining the cellular pattern of DNA methylation: DNMT1, DNMT3a, and DNMT3b. The latter two are thought to be responsible for creating DNA methylation patterns, and DNMT1 maintains established patterns through rounds of DNA replication and cell division.

As part of gene regulation, DNA methylation is coupled with demethylation, the enzyme-catalyzed removal of cytosine methyl groups. Demethylation is necessary for the epigenetic reprogramming of genes and can be passive or active. Passive demethylation is the failure to methylate new strands of DNA during replication. Active demethylation is the removal of methyl groups from methylated cytosine independent of DNA replication. DNA demethylation can occur across the entire genome (global demethylation) or may be limited to specific loci, associated with changes in gene expression.

In a recent study of 17 different human tissues, about 15 percent of CpGs, representing more than 2000 genes, were found to be hypomethylated in promoter and upstream regions in all tissues surveyed. Further analysis revealed that many of the adjacent genes were associated with housekeeping functions such as the cell cycle, transcription, and RNA processing.

The study also found that patterns of CpG hypomethylation are associated with genes involved in tissue-specific functions (**Table 19.1**). For example, in bone and joint cartilage, 11 genes associated with skeletal and cartilage development are hypomethylated and transcriptionally active, while in the bladder, 14 genes associated with muscle contraction are active. These results showed that methylome data alone were sufficient to distinguish among all the tissues studied, and that the tissues were characterized by distinctive methylation patterns that reflected their tissue-specific functions.

In some cases, tissue-specific patterns of methylation are an indication of genetic susceptibility to disease. For example, nonalcoholic fatty liver disease (NAFLD) is a major health problem in developed countries and a

TABLE 19.1 Some Tissue-Specific Gene Methylation Patterns

Tissue	Tissue-Specific Processes	Number of Genes
Artery (heart, spleen)	Blood vessel morphogenesis	12
	Angiogeneis	10
	Blood vessel development	13
Bone marrow	General cell activation	41
	White blood cell activation	33
	Immune response	62
Bone, joint cartilage	Chondrocyte differentiation	3
	Cartilage development	4
	Skeletal system development	7
Bladder	Muscle contraction	14
	Excretion	7
	Secretion	17

leading cause of chronic liver disease. Genetic factors play an important role in determining predisposition to NAFLD, but these factors explain only a small number of all cases. However, experiments using a mouse model have shown that specific patterns of DNA methylation play a primary role in susceptibility to and progression of this disease. This work further suggests that analysis of the DNA methylome can be useful in the diagnosis and treatment of NAFLD.

In mammals and other vertebrates, methylation of cytosine to form 5-methylcytosine (5mC) is the most common epigenetic modification of DNA. However, in the genomes of some eukaryotes, including the algae *Chlamydomonas reinhardtii* and the nematode *Caenorhabditis elegans*, 5mC is absent or, as in *Drosophila*, may be present at almost undetectable levels. Recent work has shown that although the methylomes of these and some other eukaryotes may not contain 5mC, they do contain adenine that has been methylated at its N6 position (6mA), a modification that may have epigenetic functions. Individually and collectively then, the evidence from initial studies shows that (a) enzymatic modification of 6mA occurs in the DNA of these organisms, (b) 6mA may be an epigenetic mark that regulates gene expression, and (c) 6mA may be involved in the transmission of epigenetic traits across generations. At this early stage, further research is needed in these species to fully explore the details of how 6mA controls gene expression. In addition, the extent to which 6mA is present in the methylomes of other organisms including mammals, and has epigenetic functions, remains to be determined.

Histone Modification and Chromatin Remodeling

Interaction of DNA with proteins that facilitate transcription is controlled by two processes: (1) chromatin remodeling, which involves the action of ATP-powered protein complexes that move, remove, or alter nucleosomes, and (2) histone modifications, which are covalent posttranslational modifications of amino acids near the N-terminal ends of histone proteins. Together, these two processes activate or repress transcription, and act as one of the primary methods of gene regulation.

Recall that chromatin is a dynamic structure composed of DNA wound around a core of eight histone proteins to form nucleosomes (Chapter 12). Normally, DNA is wound tightly around nucleosomes to form chromatin, which is further coiled and packaged to form chromosomes. In this state, regulatory regions of DNA and the genes themselves are unable to interact with proteins that facilitate transcription.

The N-terminal region of each histone extends beyond the nucleosome, forming a tail. Amino acids in these tails can be covalently modified in several ways (**Figure 19.4**).

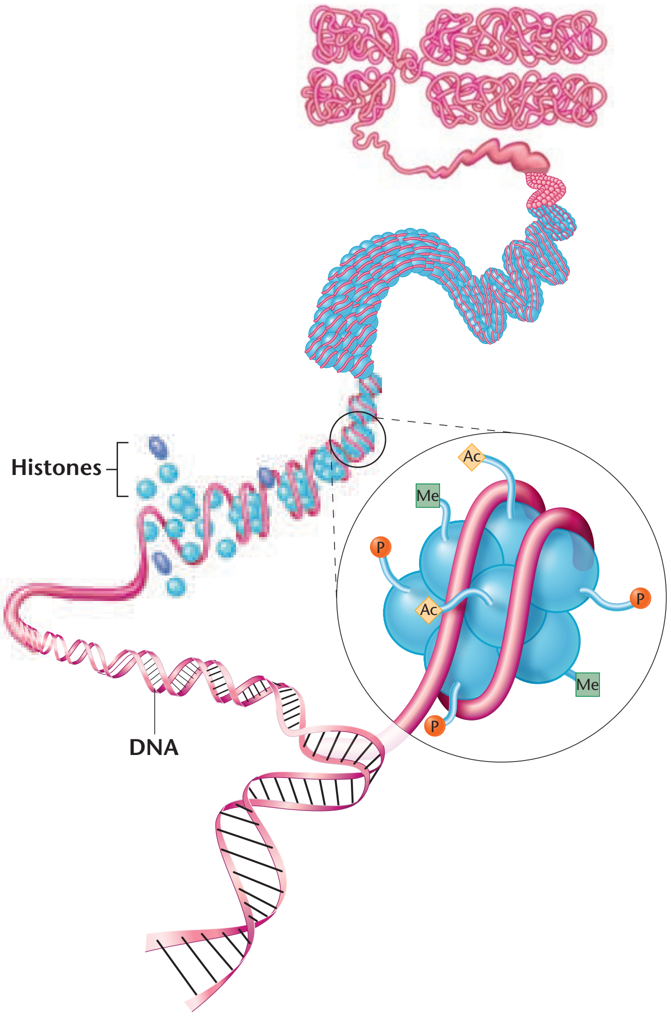


FIGURE 19.4 Histones in nucleosomes have their N-terminal tails covalently modified in epigenetic modifications that alter patterns of gene expression. Ac = acetyl group, Me = methyl group, P = phosphate group.

Several sets of proteins are involved in the process. These include proteins that add chemical groups to histones (“writers”), proteins that interpret those modifications (“readers”), and proteins that remove those chemical groups (“erasers”). Some of these histone-modifying proteins are listed in **Table 19.2**.

Over 20 different chemical modifications can be made to histones, but the major changes include the addition of acetyl, methyl, and phosphate groups (Figure 19.4). Such additions alter the structure of chromatin, making genes on nucleosomes with modified histones accessible or inaccessible for transcription. Histone acetylation, for example, relaxes the grip of histones on DNA and makes genes available for transcription [**Figure 19.5(a)**]. Furthermore, acetylation is reversible. Removing (erasing) acetyl groups contributes to changing chromatin from an “open” configuration to a “closed” state, thereby silencing genes by making them unavailable for transcription [**Figure 19.5(b)**].

TABLE 19.2 Estimated Numbers of Epigenetic Chromatin Modifier Proteins

Type	Number Identified
Writers	**78**
Protein methyltransferases	
Histone acetyltransferases	
Readers	**156**
Tudor domain-containing proteins	
MBT domain-containing proteins	
Chromodomain-containing proteins	
Erasers	**42**
Histone deacetylases	
Lysine demethylases	

Histone modifications occur at specific amino acids in the N-terminal tail of histones 2A, 2B, H3, and H4. Figure 19.6(a) shows some modifications commonly found on histones H3 and H4. Many combinations of histone modifications are possible within and between histone molecules [Figure 19.6(b)], and the sum of their complex patterns and interactions is called the **histone code**. The basic idea behind a histone code is that reversible enzymatic modification of histone amino acids (by writers and erasers) recruits nucleoplasmic proteins (readers) that either further modify chromatin structure or regulate transcription.

The code is represented in a shorthand as follows:

- Name of the histone (e.g., H3)
- Single-letter abbreviation for the amino acid (e.g., K for lysine)
- Position of the amino acid in the protein (e.g., 9)
- Type of modification (ac = acetyl, me = methyl, p = phosphate, etc.)
- Number of modifications (amino acids can be methylated one, two, or three times)

Thus, H3K27me3 represents a trimethylated lysine at position 27 from the N-terminus of histone H3.

The roles of some histone modifications in regulating gene expression are shown in Table 19.3. Specific combinations of histone modifications and interactions between modified amino acids within and between histones control the transcriptional status of a chromatin region. For example, whether or not H3K9 will be methylated is controlled by modifications made elsewhere on the protein. On one hand, if H3S10 is phosphorylated, methylation of the adjacent amino acid H3K9 is inhibited. On the other hand, if H3K14 is deacetylated, methylation of H3K9 is facilitated. Methylation of histones H3K4 and H3K36 is associated with transcriptional activation, while demethylation of H3K4, H3K9, and H3K27 is associated with gene repression.

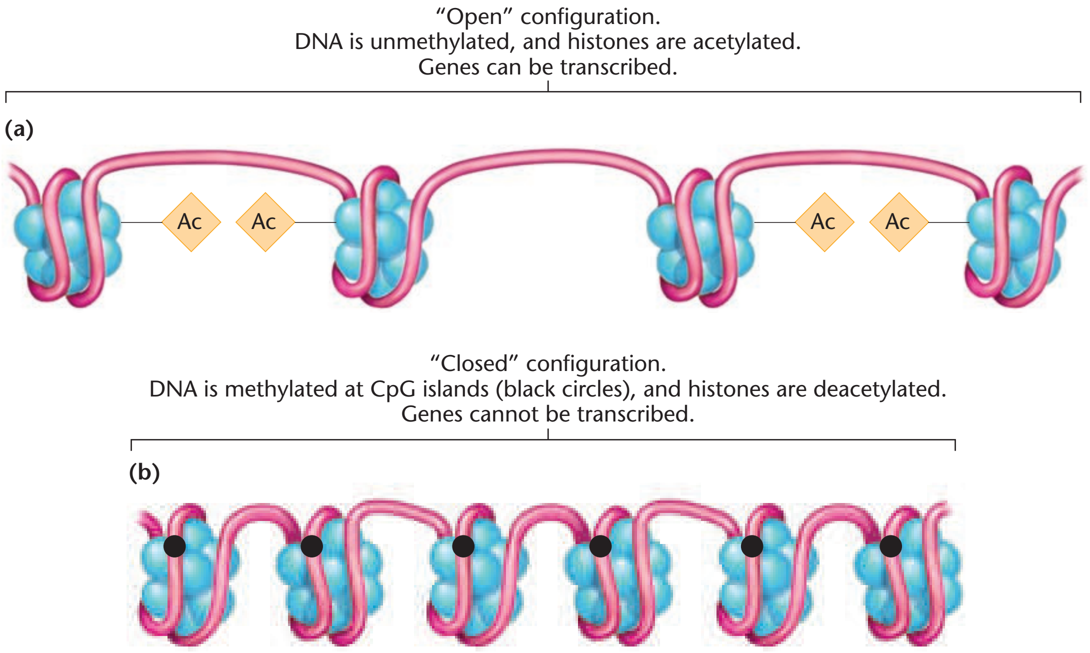

"Open" configuration.
DNA is unmethylated, and histones are acetylated.
Genes can be transcribed.
(a)
Ac
Ac
Ac
Ac
"Closed" configuration.
DNA is methylated at CpG islands (black circles), and histones are deacetylated.
Genes cannot be transcribed.
(b)

FIGURE 19.5 Epigenetic modifications to the genome alter the spacing of nucleosomes. (a) In the open configuration, nucleosome positions are shifted by chromatin remodeling, CpGs are unmethylated, and the genes on the DNA are available for transcription. (b) In the closed configuration, DNA is tightly wound onto the nucleosomes, CpGs are methylated, chemical groups have been removed from histones, and genes on the DNA are unavailable for transcription.

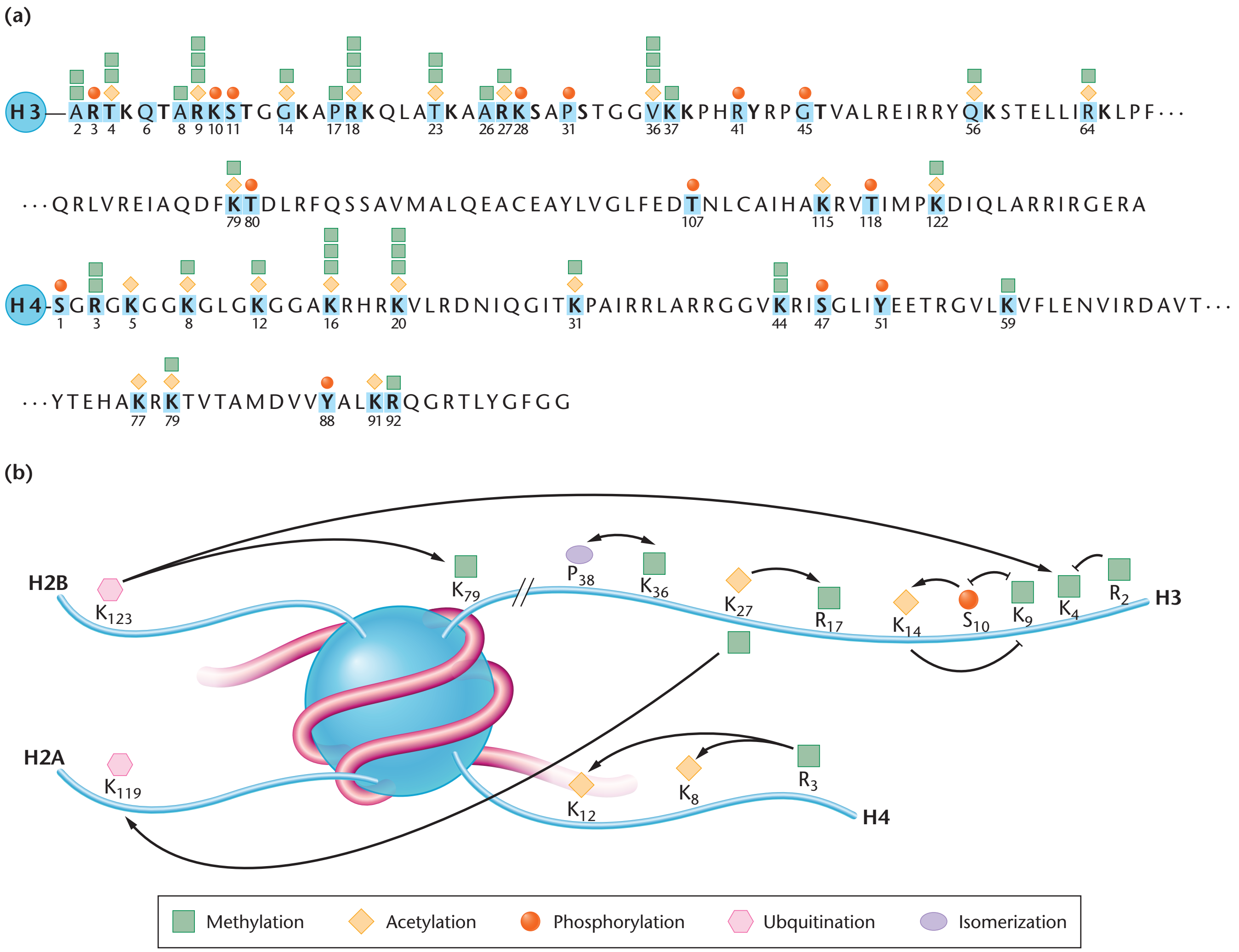


FIGURE 19.6 (a) Possible modifications within the N-terminal tails of histones H3 and H4. (All four histones—2A, 2B, H3, and H4—can be modified, but only two examples are shown here). Amino acids are represented by their one-letter abbreviations (K = lysine, R = arginine, S = serine, T = threonine, Y = tyrosine). Numbers indicate the position of the amino acid relative to the N-terminus. Symbols above the amino acids represent the chemical groups covalently attached during modification and the number of chemical groups that can be added to the amino acid (e.g., some amino acids can carry up to three methyl groups). (b) Arrows show interactions between and among modified amino acids on different histones, adding to the complexity of the histone code. A barbed arrow head indicates a positive effect; a flat arrow head indicates a negative effect. Histone modifications: ac = acetylation, iso = isomerization, me = methylation, ph = phosphorylation, ub = ubiquitination, P = proline.

TABLE 19.3 The Functions of Some Histone Modifications

Modification	H3K4	H3K9	H3S10	H3K14	H3K27	H3K36	H3K79	H4K20
Acetylation	—	Activation	—	Activation	Activation	—	—	—
First methylation (me1)	Activation	Activation	—	—	Activation	—	Activation	Activation
Second methylation (me2)	—	Repression	—	—	Repression	—	Activation	—
Third methylation (me3)	Activation	Repression	—	—	Repression	Activation	Activation Repression	—
Phosphorylation	—	—	Repression	—	—	—	—	—

The histone code can be extremely complex. Considering only the addition of one, two, or three methyl groups to amino acids in H3, there are about 280 billion combinations. When all possible modifications of all histones are considered, the number of possible combinations is truly astronomical. Much work lies ahead to identify all their epigenetic roles.

What we do know is that this wealth of possible combinations allows differentiated cells to carry out cell-specific patterns of gene transcription and to respond to external signals that modify these patterns *without any changes in DNA sequence*.

Now Solve This

19.1 Although histone modifications can activate or silence genes, these covalent alterations are made to protein molecules involved in nucleosome structure and not to the DNA carrying the activated or silenced allele. If the fixed pattern of active and silenced alleles is to be carried through multiple cell divisions, would you expect the histone modifications to be in *cis* or *trans* to the affected alleles? Why?

- **Hint:** *This problem involves an understanding of how many copies of each histone are contained in a nucleosome and of the spatial relationship between the histones and the DNA wound around the nucleosome.*

Short and Long Noncoding RNAs

In addition to messenger RNA (mRNA), genome transcription produces several classes of **noncoding RNAs (ncRNAs)** which are transcribed from DNA, but not translated into proteins. The ncRNAs related to epigenetic regulation include two groups: (1) short ncRNAs (less than 31 nucleotides) and (2) long ncRNAs (greater than 200 nucleotides). Both types of ncRNAs have several roles, including the formation of heterochromatin, histone modification, site-specific DNA methylation, and gene silencing. They are also important in epigenetic regulatory networks.

There are three classes of short ncRNAs: miRNAs (microRNAs), siRNAs (short interfering RNAs), and piRNAs (piwi-interacting RNAs). miRNAs and siRNAs are transcribed as precursor molecules about 70–100 nucleotides long that contain a double-stranded stem-loop and single-stranded regions. After several processing steps that shorten the RNAs to lengths of 20–25 ribonucleotides, these RNAs act as repressors of gene expression (see Chapter 18 for a detailed discussion of these short ncRNAs). The origin of piRNAs is unclear, but they interact with proteins to form RNA-protein complexes that participate in epigenetic gene silencing in germ cells.

Long noncoding RNAs (lncRNAs) share many properties with mRNAs; they often have 5′ caps, 3′ poly-A tails, and are spliced. What distinguishes lncRNAs from coding (mRNA) transcripts is the lack of an extended open reading frame that codes for the insertion of amino acids into a polypeptide.

The discovery of lncRNAs was a by-product of the Human Genome Project. Genomic sequencing identified several thousand RNA genes that were not protein coding. More recent studies using RNA sequencing have identified over 14,000 lncRNA genes in the human genome.

lncRNA loci are often classified by their relationship to nearby protein-coding genes. *Antisense lncRNA genes* partially overlap protein-coding genes and are transcribed in the opposite direction to the protein-coding gene [Figure 19.7(a)]. *Intronic lncRNA genes* are located within introns, and their transcription does not overlap with the adjacent exons of protein-coding genes [Figure 19.7(b)]. *Bidirectional lncRNA genes* use the promoter of a protein-coding gene but are transcribed in the opposite direction [Figure 19.7(c)]. *Intergenic lncRNA genes* are discrete transcription units located outside protein-coding genes [Figure 19.7(d)].

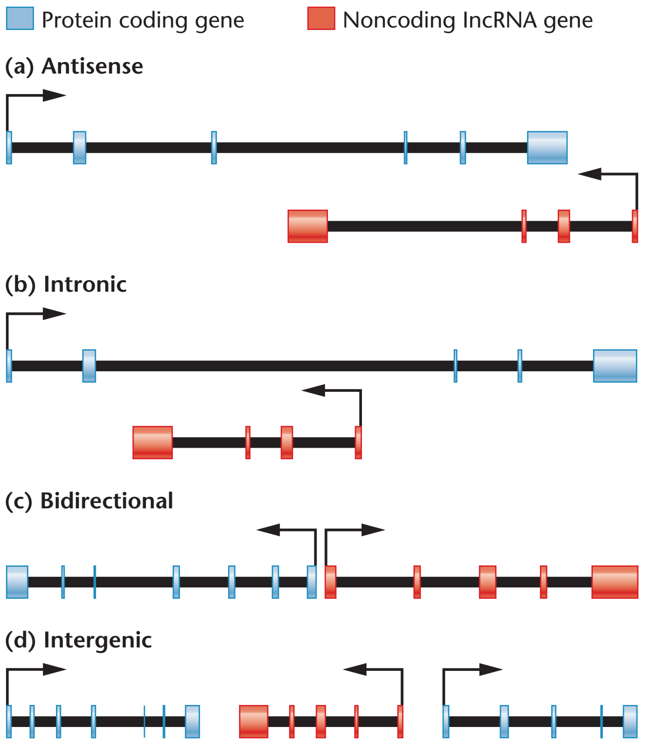


FIGURE 19.7 Four classes of lncRNA loci. (a) *Antisense lncRNA genes* are usually partially 3′ to a protein-coding gene, have their own promoter, and are transcribed in the opposite direction. (b) *Intronic lncRNA genes* are contained in introns within protein-coding genes. (c) *Bidirectional lncRNA genes* have no sequence overlap with protein-coding genes, but use the promoter of an adjacent protein-coding gene and transcribe in the opposite direction. (d) *Intergenic lncRNA genes* are completely independent loci, do not overlap with protein-coding genes, and use their own promoters.

In spite of the large number of lncRNA loci and differences in gene organization and transcriptional pattern, it is clear that lncRNAs share some common properties: (1) they form RNA-protein complexes with many different chromatin regulators, (2) they deliver these complexes to specific locations in the genome, and (3) they participate in chromatin remodeling, interact with transcription factors, and carry out other, as yet unidentified mechanisms of gene regulation.

Research to date suggests that lncRNAs can interact with selected regions of the genome by several mechanisms. First, lncRNAs can act as *decoys* to bind transcription factors and prevent them from interacting with target genes [**Figure 19.8(a)**]. lncRNAs can also serve as platforms or *adapters*, to bring two or more proteins together to form a functional complex [**Figure 19.8(b)**]. For example, the lncRNA HOTAIR binds two protein complexes that allow their coordinated action to methylate H3K27 and demethylate H3K4me2, a combination of histone modifications that silences target genes.

lncRNAs can also target chromatin remodeling enzyme complexes that are involved in gene silencing [**Figure 19.8(c)**]. In this capacity, the lncRNAs serve as *guides* to target gene silencing in an allele-specific fashion, highlighting two of the basic functions of lncRNAs. More recently, it has been proposed that lncRNA guidance occurs through another mechanism via the looping of chromosomal regions in a model similar to the way *enhancers* bring proteins to the upstream regulatory sequences of a gene [**Figure 19.8(d)**].

lncRNAs are found in the nucleus and the cytoplasm and, through a variety of mechanisms, are involved in both transcriptional and posttranscriptional regulation of gene expression. As epigenetic initiators, lncRNAs bind to chromatin-modifying enzymes and direct their activity to specific regions of the genome. At these sites, the lncRNAs direct chromatin modification, altering the pattern of gene expression.

In summary, epigenetic modifications alter chromatin structure by several mechanisms: DNA methylation, reversible covalent modification of histones, and action of short and long RNAs, all without changing the sequence of genomic DNA. This suite of epigenetic changes creates an epigenome that, in turn, can regulate normal development and generate changes in gene expression as a response to environmental signals.

19.2 Epigenetics and Monoallelic Gene Expression

Mammals inherit a maternal and a paternal copy of each gene, and aside from genes on the inactivated X chromosome in females, both copies of these genes are usually expressed at equal levels in the offspring. However, in some cases only one allele is transcribed, while the other allele is transcriptionally silent. This phenomenon is called **monoallelic expression (MAE)**.

There are three major classes of MAE. One is **genomic imprinting**, in which genes are expressed in a *parent-of-origin pattern*; that is, certain genes show expression of only the maternal allele or the paternal allele. The remaining two classes fall into the category of *random* monoallelic expression. First is the random inactivation of one X chromosome in the cells of mammalian females, which compensates for their increased dosage of X-linked genes (recall that mammalian males have only one X chromosome). Second, a randomly generated pattern of allele inactivation is observed in a significant number of autosomal genes. The pattern in autosomes also occurs independent of parental origin. In mice and humans, studies of several thousand genes in numerous cell types show that this form of gene regulation was found in 10 to 20 percent of these genes. We will look at each of these three classes in turn.

Parent-of-Origin Monoallelic Expression: Imprinting

Parentally imprinted genes are marked in male and female germ-line cells during gamete formation; the fertilized egg

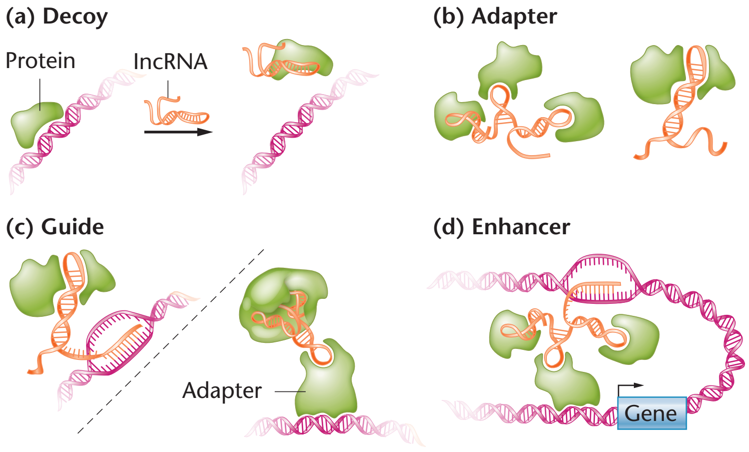


FIGURE 19.8 Four models for lncRNA mechanisms of action. (a) *The decoy model.* lncRNA-binding sites compete with transcription initiation sites on protein-coding genes to prevent transcription. (b) *The adapter model.* lncRNAs act as a platform or adapter for two or more proteins to form active DNA-binding protein complexes. (c) *The guide model.* lncRNAs recruit protein complexes (e.g., chromatin modifiers) and guide them to specific loci. This interaction can be via RNA–DNA or RNA–protein interactions. (d) *The enhancer model.* lncRNAs bind to DNA regions upstream or downstream of genes forming enhancer-like loops to regulate gene action.

thus has different marks on the copies of certain genes that came from the mother or the father. How is this marking accomplished?

To begin with, the DNA carried by sperm and eggs are highly methylated. However, shortly after fertilization, most of the germ-line methylation marks are erased. This modification of the DNA resets and provides embryonic cells with a clean pre-epigenetic state, so to speak, allowing them to undergo new epigenetic modifications to form the more than 200 cell types found in the adult body. About the same time the embryo is implanting in the wall of the uterus, cells take on tissue-specific epigenetic identities, and methylation patterns and histone modifications change rapidly to reflect those seen in differentiated cells.

Some genomic regions, however, escape these rounds of global demethylation and remethylation. The genes contained in these regions remain imprinted with the methylation marks of the maternal and/or paternal chromosomes. These original parental patterns of methylation produce allele-specific imprinting. Imprinted alleles remain transcriptionally silent during embryogenesis and later stages of development. For example, if the allele inherited from the father is imprinted, it is silenced, and only the allele from the mother is expressed.

In humans, imprinted genes are usually found in clusters on the same chromosome that can occupy more than 1000 kb of DNA. Because these genes are located near each other at a limited number of sites in the genome, mutation in one imprinted gene can often affect the function of adjacent or coordinately controlled imprinted genes, thereby amplifying the mutation's phenotypic impact. These mutations in imprinted genes can arise through changes in the DNA sequence or by resultant dysfunctional epigenetic changes, called **epimutations**, both of which can cause heritable changes in gene activity.

Occasionally, the imprinting process goes awry and is described as being dysfunctional. In such cases, the imprinting defects can cause human disorders such as Beckwith–Wiedemann syndrome, Prader–Willi syndrome, Angelman syndrome, and several other diseases (**Table 19.4**). However, given the number of imprinting-susceptible candidate genes and the possibility that additional imprinted genes still remain to be discovered, the overall number of imprinting-related genetic disorders may be quite high.

TABLE 19.4 Some Imprinting Disorders in Humans

Disorder	Locus
Albright hereditary osteodystrophy	20q13
Angelman syndrome	15q11–q15
Beckwith–Wiedemann syndrome	11p15
Prader–Willi syndrome	15q11–q15
Silver–Russell syndrome	Chromosome 7
Uniparental disomy 14	Chromosome 14

In humans, most known imprinted genes encode growth factors or other growth-regulating genes. An autosomal dominant disorder associated with imprinting, Beckwith–Wiedemann syndrome (BWS) occurs in about 1 in 13,700 births and offers insight into how disruptions of epigenetic imprinting can lead to an abnormal phenotype. BWS is a prenatal overgrowth disorder typified by abdominal wall defects, enlarged organs, high birth weight, and a predisposition to cancer. BWS is not caused by mutation in the DNA sequence of the gene, nor is it associated with any chromosomal aberrations. The genes associated with BWS are located in a cluster of epigenetically imprinted genes on the short arm of chromosome 11 (**Figure 19.9**). BWS is a disorder of imprinting and is caused by abnormal patterns of DNA methylation resulting in altered patterns of gene expression.

All the genes in this particular cluster are known to regulate growth during prenatal development. Two closely linked genes in this cluster are *insulin-like growth factor 2* (*IGF2*), whose encoded protein plays an important role in growth and development, and *H19*, which is transcribed into an ncRNA. These two genes are separated by an imprinting control region (ICR), which controls the expression of both genes. Normally, the ICR on the paternal copy of chromosome 11 is methylated, allowing expression of the paternal *IGF2* allele but maintaining the paternal *H19* allele in a silenced state [**Figure 19.9(a)**].

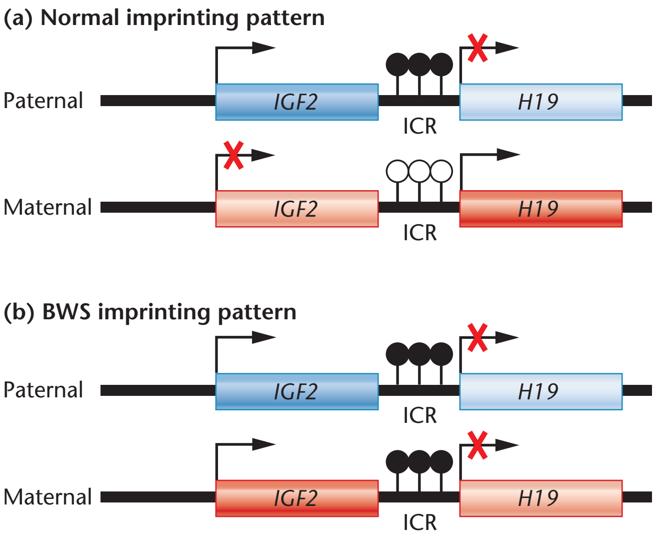


FIGURE 19.9 The imprinted region of human chromosome 11. (a) In normal imprinting, the ICR on the paternal chromosome is methylated (filled circles); the *IGF2* allele is active and the *H19* allele is silent. The ICR on the maternal chromosome is not methylated (open circles), and the *IGF2* allele is silent while the *H19* allele is active. (b) In one form of BWS, both the maternal and paternal ICRs are methylated (filled circles), both *IGF2* alleles are active, and both *H19* alleles are silent. The result is dysregulation of cell growth, resulting in the overgrowth of structures that are characteristic of BWS.

Reciprocally, on the maternal copy of chromosome 11, the ICR is unmethylated allowing for the expression of the maternal *H19* allele, while the maternal *IGF2* allele is maintained in a silenced state.

In BWS, several mechanisms can cause dysregulation of growth. In one form of BWS, both copies of the ICR are methylated, and both the maternal and paternal *IGF2* alleles are transcriptionally active, resulting in the overgrowth of tissues that are characteristic of this disease. In this situation, transcription of both *IGF2* alleles is accompanied by silencing of both copies of the *H19* allele, further compounding the overgrowth of tissues [**Figure 19.9(b)**].

Prader–Willi syndrome (PWS) and Angelman syndrome (AS) were the first examples of imprinting disorders in humans and are caused by differential parental imprinting of the same region on the long arm of chromosome 15, or by the deletion of part of this region. The genes involved in these two disorders are part of a cluster of epigenetically imprinted genes expressed from either the paternal or maternal copies of chromosome 15, but not both. PWS occurs in about 1 in 15,000 births and causes intellectual disability, an uncontrollable appetite, obesity, diabetes, and growth disorders. In normal individuals, the genes associated with PWS are paternally expressed and maternally silenced. These genes are directly or indirectly involved in the development and function of the brain. The epigenetically controlled disruption of the imprinting pattern results in silencing of the paternal alleles of these genes, resulting in PWS.

AS primarily affects the nervous system; affected individuals have intellectual disabilities, along with involuntary muscle contractions (chorea), and seizures. This disorder occurs in about 1 in 24,000 births.

AS is caused by abnormal imprinting of a single gene (*UBE3A*). Normally, both the maternal and paternal alleles are expressed in many tissues of the body. In some parts of the brain, the paternal copy is typically imprinted and silenced, leaving only the maternal copy to be expressed. Thus, inactivation or loss of the maternal copy of the *UBE3A* gene means there are no active copies of the gene in some parts of the brain, and the result is AS.

The known number of imprinted genes represents only a small fraction (less than 1 percent) of the mammalian genome, but they play major roles in regulating growth and development during the prenatal stage. Because they act so early in life, any external or internal factors that disturb the epigenetic patterns of imprinting or the expression of these imprinted genes can have serious phenotypic consequences.

Random Monoallelic Expression: Inactivation of the X Chromosome

The random inactivation of an X chromosome in cells of female mammals was actually the first example of epigenetic allele-specific regulation to be identified. At an early stage of development, about half of embryonic cells randomly inactivate the maternal X chromosome and the other half inactivate the paternal X chromosome, effectively silencing almost all the 900 or so genes on whichever homolog is inactivated. Once inactivated, the same X chromosome remains silenced in all cells descended from this progenitor cell.

How does X inactivation occur? Several lncRNAs play a key role in this process. Two of the major contributors are Xist (X inactive specific transcript), and Tsix (Xist spelled backward), which are sense and antisense transcripts of the same gene (transcribed in opposite directions). The Xist lncRNA is expressed on the inactivated X chromosome and coats the entire chromosome, converting it into a Barr body (see Chapter 7), which is a highly condensed and genetically silent chromatin structure. The lncRNA Tsix is expressed on the active X chromosome and represses expression of the Xist lncRNA, thus preventing the active X chromosome from being silenced.

Random Monoallelic Expression of Autosomal Genes

Monoallelic expression (MAE) of autosomal genes has been known for decades in the immunoglobulin family and in olfactory receptors, but the phenomenon was thought to be limited to a small number of gene families. Recently however, genome-wide analysis of allele-specific expression in mice and humans led to the surprising discovery that MAE is a widespread event, involving 10 to 20 percent of autosomal genes in a range of different cell types.

Unlike imprinted genes, which are located in clusters, autosomal MAE genes are scattered throughout the genome. Because autosomal MAE is a random process, four states of expression for a gene are possible in cells of a given tissue: (1) expression of both alleles (biallelic expression), (2) expression of only the maternal allele, (3) expression of only the paternal allele, or (4) expression of neither allele (**Figure 19.10**). This generates a tissue-specific spectrum of biallelic, maternal, paternal, or no expression of a gene.

These different patterns of expression, all present in the same tissue, can have an impact on the phenotype and may offer a molecular explanation for the incomplete penetrance of traits observed in some genetic disorders. (See Chapter 4 for a discussion of incomplete penetrance.)

The first step in generating autosomal monoallelic expression is *stochastic*, meaning that it is a cell-specific, random event, paralleling the random selection of X chromosomes for inactivation in dosage compensation. This event may involve silencing one allele in a cell with biallelic expression, or activating an allele from a silent gene. Feedback mechanisms may prevent activation or silencing of the second allele. The allele choice may occur early in development and is transmitted by cell division, resulting in the mosaic clonal structure of tissues.

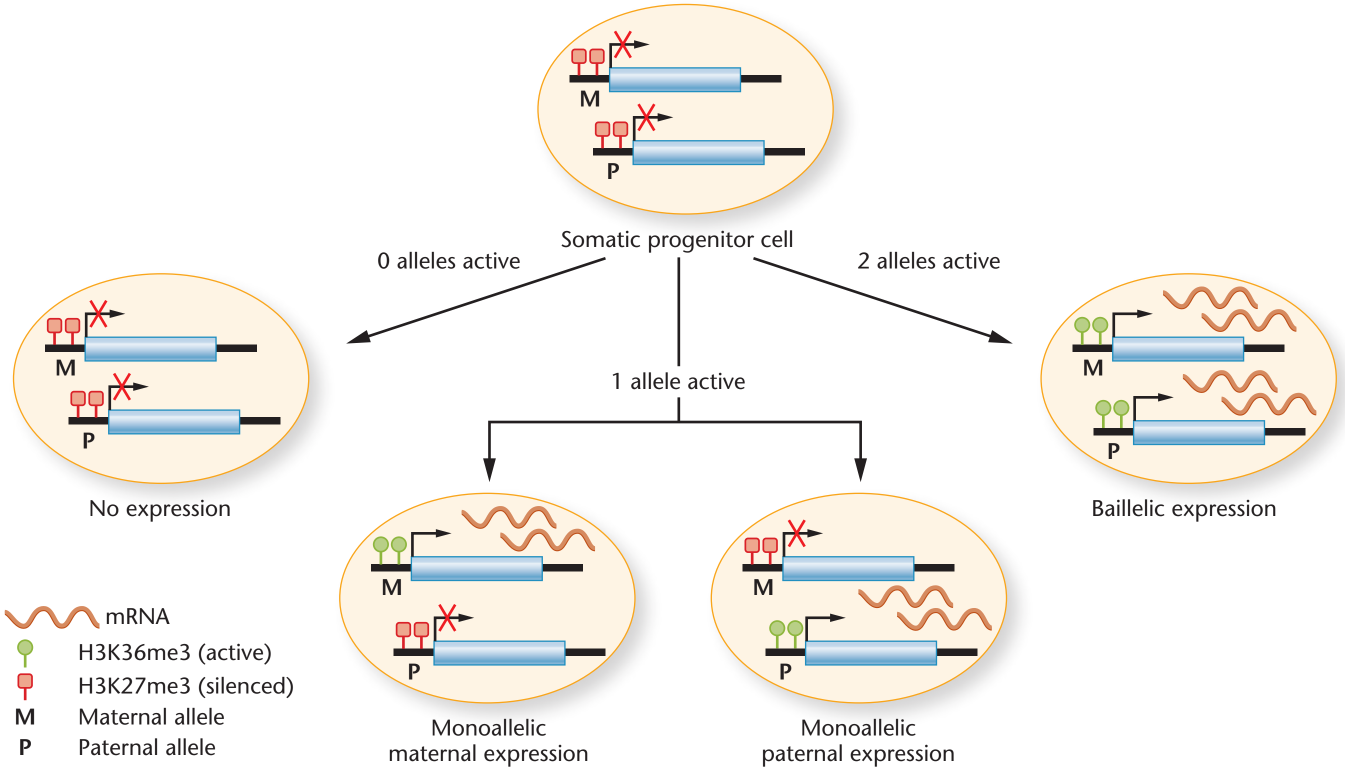


FIGURE 19.10 A model of randomly determined epigenetic gene regulation. Early in development, beginning with the loss of pluripotency and the early stages of cell determination in somatic progenitor cells, random monoallelic regulation of homologous alleles on a cell-by-cell basis results in a population of cells within a tissue expressing combinations of allele expression: neither allele, only maternal, only paternal, or both alleles (biallelic expression). Once established, these regulatory states are passed on in clonal fashion, resulting in a tissue with a mosaic pattern of expression for a specific gene.

Using ChIP-sequencing, a method that analyzes protein-DNA interactions, along with RNA-seq methods (both methods are discussed in Chapter 21), researchers have identified a distinct histone modification signature associated with MAE. By analyzing a number of epigenetic marks present in a wide range of cell types, they established that two modifications, H3K27me3 and H3K36me3, explain most of the difference between cells with monoallelic expression of a given gene and cells with biallelic expression of the same gene. In MAE cells, the H3K27me3 marker, associated with gene silencing, is linked to the inactive allele, while the H3K36me3 marker, associated with transcription, is linked to the active allele. This chromatin signature is a powerful and reliable predictor of MAE activity in many cell types and offers a way of exploring the relationship between epigenetics and disease.

Assisted Reproductive Technologies (ART) and Imprinting Defects

In the United States, **assisted reproductive technologies (ART)**, including *in vitro* fertilization (IVF), are now used in over 1 percent of all births. Over the past decade, several studies have suggested that children born through the use of ART have an increased risk for imprinting errors (epimutations) caused by the manipulation of gametes or embryos.

For example, the use of ART results in a four- to nine-fold increased risk of Beckwith–Wiedemann syndrome (BWS); in addition, there are increased risks for Prader–Willi syndrome (PWS) and Angelman syndrome (AS). Studies of children with BWS or AS conceived by IVF have shown that they have reduced levels or complete loss of maternal-specific methylation at known imprinting sites in the genome, confirming the role of epigenetics in these cases. ART procedures are done at times when the oocyte and the early embryo are undergoing epigenetic reprogramming. It appears that disturbances in epigenetic programming at sensitive times during development are responsible for the increased risk of these disorders.

Although imprinting errors are uncommon in the general population (BWS occurs in only about 1 in 13,700 births), epimutations may be a significant risk factor for those conceived by ART.

19.3 Epigenetics and Cancer

Cancer cells have mutations in many different genes, both those that clearly contribute to cancer development and those that contribute to suppression of cancer, as well as mutations in genes that may not directly contribute to cancer. Following the discovery of cancer-associated genes, including those that promote cell division (proto-oncogenes) or inhibit cell division (tumor-suppressor genes), research into the genetics of cancer has revealed mutations in genes with many different functions associated with cancer progression, including DNA repair, apoptosis, genome stability, and cell signaling. (For a full discussion of cancer genetics, see Chapter 24.)

Originally it was thought that cancer is clonal in origin and begins in a single cell that has accumulated a suite of mutations that allow it to escape control of the cell cycle. Subsequent mutations allow cells of the tumor to become metastatic, spreading the cancer to other locations in the body where new malignant tumors appear. However, converging lines of evidence are now clarifying the importance of epigenetic changes in the initiation and maintenance of malignancy. These findings are helping researchers understand the properties of cancer cells that are difficult to explain by the action of mutant alleles alone. Evidence for the role of epigenetic changes in cancer has established epigenomic changes as a major pathway for the formation and spread of malignant cells.

DNA Methylation and Cancer

As far back as the 1980s, researchers observed that cancer cells had much lower levels of methylation than normal cells derived from the same tissue. Subsequent research by many investigators showed that complex changes in DNA methylation patterns are associated with cancer. These studies showed that genomic hypomethylation is a property of all cancers examined to date.

DNA hypomethylation reverses the silencing of genes, leading to unrestricted transcription of many gene sets—including those associated with the development of cancer. It also relaxes control over imprinted genes, causing cells to acquire new growth properties. Hypomethylation of repetitive DNA sequences in heterochromatic regions increases chromosome rearrangements and changes in chromosome number, both of which are characteristic of cancer cells. In addition, hypomethylation of repetitive sequences leads to transcriptional activation of transposable DNA sequences such as LINEs and SINEs, further increasing genomic instability.

Despite the fact that cancer cells are characterized by global hypomethylation, selected regions of their genome are hypermethylated when compared to normal cells. Selective hypermethylation of promoter-associated CpG islands silences certain genes, including tumor-suppressor genes, often in a tumor-specific fashion (**Table 19.5**). Analysis of these patterns, called the CpG island methylator phenotype (CIMP), provides a way to identify tumor types and subtypes and predict the sites to which the tumor may metastasize.

For example, the promoter region of the breast cancer gene *BRCA1* is hypomethylated in normal cells, but is hypermethylated and inactivated in many cases of breast and ovarian cancer. In another example, silencing of the DNA repair gene *MLH1* by hypermethylation is a key step in the development of some forms of colon cancer. *MLH1* illustrates how epimutations can be involved in tumor formation, either alone or in combination with genetic changes (**Figure 19.11**).

In point of fact, cancer is now viewed as a disease that usually results from the accumulation of both genetic *and* epigenetic changes (**Figure 19.12**). For example, in a bladder cancer cell line, one allele of a tumor-suppressor gene, *CDKN2A,* is mutated, and the other, normal, allele is silenced by hypermethylation. Because both alleles are inactivated (although by different mechanisms), cells are able to escape control of the cell cycle and divide continuously. Even more striking, in ovarian cancer, mutations in nine specific genes are predominant, but promoter hypermethylation is observed in 168 genes. These genes are epigenetically silenced, and their reduced expression is linked to the development and maintenance of this cancer.

The broad pattern of hypermethylation seen in cancer cells and the many functions of the affected genes suggest that this phenomenon may result from a widespread

TABLE 19.5 Some Human Cancer-Related Genes Inactivated by Hypermethylation

Gene	Locus	Function	Related Cancers
BRCA1	17q21	DNA repair	Breast, ovarian
APC	5q21	Nucleocytoplasmic signaling	Colorectal, duodenal
MLH1	3p21	DNA repair	Colon, stomach
RB1	13q14	Cell-cycle control point	Retinoblastoma, osteosarcoma
AR	Xq11–12	Nuclear receptor for androgen; transcriptional activator	Prostate
ESR1	6q25	Nuclear receptor for estrogen; transcriptional activator	Breast, colorectal

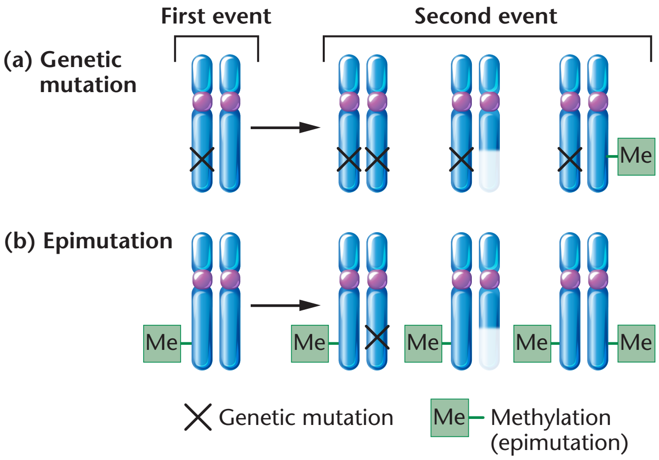


FIGURE 19.11 The role of epimutations versus genetic mutations in the initiation of cancer. (a) An inherited genetic mutation causes the loss of a tumor-suppressor allele. Several mechanisms can cause the loss or silencing of the second allele: mutation, chromosomal aberration, or an epimutation. (b) An epimutation silences one allele of a tumor-suppressor gene. The second allele can be lost through genetic mutation, chromosomal aberration, or silencing by an epigenetic event.

deregulation of the methylation process rather than a targeted event.

In fact, many of the mechanisms that cause epigenetic changes in cancer cells are not well understood, partly because the changes take place very early in the conversion of a normal cell to a cancerous one, and partly because by the time the cancer is detected, alterations in methylation patterns have already occurred. The DNA repair gene *MLH1*, for example, plays an important role in genome stability, and silencing this gene by hypermethylation (as described on the previous page) causes instability in repetitive microsatellite sequences, which, in turn, is an important step in the development of colon cancer and several other cancers. In some individuals with colon cancer, the *MLH1* promoter in normal cells of the colon is already silenced by hypermethylation, indicating that this epigenetic event occurs very early in tumor formation, before the development of downstream genetic mutations.

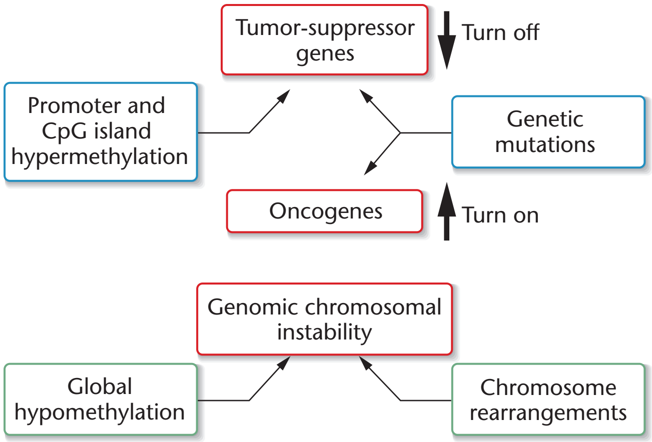

Tumor-suppressor genes
Turn off
Promoter and CpG island hypermethylation
Genetic mutations
Oncogenes
Turn on
Genomic chromosomal instability
Global hypomethylation
Chromosome rearrangements

FIGURE 19.12 The development and maintenance of malignant growth in cancer involves the interaction of gene mutations, hypomethylation, hypermethylation, overexpression of oncogenes, and the silencing of tumor-suppressor genes.

In sum, several lines of evidence support the role of epigenetic alterations in cancer:

1. Global hypomethylation may cause genomic instability and the large-scale chromosomal changes that are a characteristic feature of cancer.

2. Epigenetic mechanisms can replace mutations as a way of silencing individual tumor-suppressor genes or activating oncogenes.

3. Epigenetic modifications can silence multiple genes, making them more effective in transforming normal cells into malignant cells than sequential mutations of single genes.

Chromatin Remodeling and Histone Modification in Cancer

Chromatin remodeling involves two epigenetic systems: (1) the action of chromatin remodeling complexes to move, remove, or restructure nucleosomes to activate or silence transcription; and (2) covalent modification of histones to reversibly regulate transcription. Mutations in components of chromatin remodeling complexes and the histone modification system allow cells to escape cell-cycle control and divide continuously.

In addition to abnormal regulation of methylation, many cancers also have altered patterns of chromatin remodeling. One form of remodeling is controlled by the reversible covalent modification of histone proteins in nucleosome cores. Recall from Section 19.1 that this process involves three classes of enzymes: *writers* that add chemical groups (such as acetyls) to histones; *erasers* that remove these groups; and *readers* that recognize and read the epigenetic marks. Abnormal regulation of each of these enzyme classes results in disrupted histone profiles and is associated with a variety of cancer subtypes.

Histone acetylation is strongly correlated with activation of transcription (Chapter 17). Acetylation reduces the strength of histone interaction with DNA, making promoters available to transcription factors and RNA polymerase. Mutations in genes of both the histone acetyltransferase (HAT) family, which encode enzymes that add acetyl groups, and genes of the histone deacetylase (HDAC) family, which encode enzymes that remove acetyl groups and induce gene silencing, are linked to the development of cancer. For example, individuals with Rubinstein–Taybi syndrome have a newly arising mutation that produces a dysfunctional HAT enzyme, and about 5 percent of all cases develop cancer as a result of silencing tumor-suppressor genes. Abnormalities in histone deacetylation have been identified as an early event in the transformation of normal cells into cancer cells. HDAC complexes are selectively recruited to tumor-suppressor genes by mutated, oncogenic DNA-binding proteins. Action of the HDAC complexes at these genes converts the chromatin to a closed configuration and inhibits transcription, causing the cell to gain a growth advantage and move closer to becoming a malignant cell characterized by uncontrolled division.

More recently, research using whole exome sequencing (WES) (Chapter 21) has identified high frequencies of mutations of some protein components of chromatin remodeling complexes in human cancers. Most of the mutations have been reported in subunits of the SWI/SNF complex (Chapter 17) and found in about 20 percent of all cancers examined, covering a wide range of tumor types. Because there are a large number of different proteins in this complex, specific subunit mutations are associated with specific cancers. Mutations within tumor cells can be homozygous or, in most cases, heterozygous, making them dosage sensitive.

In summary, several lines of evidence support the role of epigenetic alterations in cancer: (1) epigenetic mechanisms can replace mutations as a way of silencing individual tumor-suppressor genes or activating oncogenes; (2) global hypomethylation may cause genomic instability and the large-scale chromosomal changes that are a characteristic feature of cancer; and (3) epigenetic modifications can silence multiple genes, making them more effective in transforming normal cells into malignant cells than sequential mutations of single genes.

Epigenetic Cancer Therapy

The fact that unlike genetic alterations, which are almost impossible to reverse, epigenetic changes are potentially reversible has inspired researchers to look for new classes of drugs to treat cancer. Epigenetic cancer therapy is focused on reprogramming gene expression through the use of drugs that alter events in chromatin remodeling in order to change the pattern of gene expression from malignant to normal.

The focus of epigenetic therapy in the development of first-generation drugs has been the reactivation of genes silenced by methylation or histone modification, essentially reprogramming the pattern of gene expression in cancer cells.

Several epigenetic drugs have been approved by the U.S. Food and Drug Administration, and another 18 or more drugs are in clinical trials. One approved drug, Vidaza (azacytidine), is used in the treatment of myelodysplastic syndrome, a precursor to leukemia, and acute myeloid leukemia. This drug is an analog of cytidine and is incorporated into DNA during replication during the S phase of the cell cycle. Methylation enzymes (methyltransferases) bind irreversibly to this analog, preventing methylation of DNA at many other sites, effectively reducing the amount of methylation in cancer cells.

Other drugs that inhibit histone deacetylases (HDACs) have been approved by the FDA for use in epigenetic therapy. Laboratory experiments with cancer cell lines indicate that inhibiting HDAC activity results in the re-expression of tumor-suppressor genes. HDAC inhibitors like Zolinza (vorinostat) are used to treat certain forms of lymphoma.

The development of epigenetic drugs for cancer therapy is still in its infancy. The approved epigenetic drugs are only moderately effective on their own and are best used in combination with other anticancer drugs. To develop more effective drugs, several important questions remain to be answered: What causes cancer cells to respond to certain epigenetic drugs? Which combinations of chromatin remodeling drugs, histone modification drugs, and conventional anticancer drugs are most effective on specific cancers? Which epigenetic markers will be effective in predicting sensitivity or resistance to newly developed drugs? Further research into the mechanisms and locations of epigenetic genome modification in cancer cells will allow the design of more potent drugs to target epigenetic events as a form of cancer therapy.

19.4 Epigenetic Traits Are Heritable

The discovery that DNA is the cellular molecule that contains genetic information encoded in its nucleotide sequence was regarded as the ultimate explanation for Mendel's laws of inheritance. Random variation of this sequence produced during DNA replication or by external agents via mutation, is responsible for generating new alleles, which in turn, are acted on by natural selection.

In contrast, however, cases of inheritance that are not based on alterations in DNA sequence have been described in many different organisms, including bacteria, plants, and animals, including humans. These phenomena challenge the idea that variations in DNA sequence are the sole link between genotype and phenotype. The elucidation of epigenetic mechanisms that modify chromatin and make non-sequence-based changes to DNA offer an explanation of how environmental effects can alter the genome and lead to small- as well as large-scale changes in patterns of gene expression.

Environmental Induction of Epigenetic Change

Environmental agents including nutrition, exposure to chemicals, and physical factors, such as temperature, can alter gene expression by affecting the epigenome. In humans it is difficult to determine the relative contributions of the environment as factors in altering the epigenome, but there is indirect evidence that changes in nutrition and exposure to agents that affect a developing fetus can have detrimental effects during adulthood.

During World War II, a famine in the western part of the Netherlands lasted from November 1944 to May 1945. During this time, daily food intake for adults was limited to 400–800 calories, well below the normal levels of 1800–2000 calories. In the period immediately after the famine, mortality rates in this population doubled, and most of this increase was attributed to malnutrition. Studies were conducted for decades afterward on the health of adult children of women who were pregnant or became pregnant during the famine. Overall, the findings show that the severity of health effects was correlated with prenatal time of exposure to famine conditions. Adults who were exposed early in prenatal development (an F_1 generation) had higher rates of several disorders—including obesity, heart disease, and breast cancer—and higher mortality rates than adults exposed later in development. In addition, as adults, there was increased risk for schizophrenia and other neuropsychiatric disorders for those with early exposure, perhaps related to nutritional deficiencies during development of the brain and nervous system. Some effects persisted in the F_2 generation, where adults had abnormal patterns of growth and increased rates of obesity. Other studies in China and Africa on the adult children of women who were pregnant or became pregnant during times of famine confirm the deleterious impact of poor maternal nutrition during pregnancy on their offspring and subsequent generations.

More direct evidence for the role of environmental factors in modifying the epigenome comes from studies in experimental animals. A low-protein diet fed to pregnant rats causes changes in the expression of several genes in both the F_1 and F_2 offspring. Increased expression of these genes is associated with hypomethylation of their promoters. Other evidence indicates that epigenetic changes triggered by this diet modification were gene specific.

Another dramatic example of how epigenome modifications affect the phenotype comes from the study of coat color in mice, where color is controlled by the dominant allele *Agouti* (*A*). In homozygous *AA* mice, the allele is active only during a specific time during hair development, resulting in a yellow band on an otherwise black hair shaft, producing the agouti phenotype. A nonlethal mutant allele (A^{vy}) causes yellow pigment formation along the entire hair shaft, resulting in yellow fur color. This allele is the result of the insertion of a transposable element near the transcription start site of the *Agouti* gene. A promoter element within the transposon is responsible for this change in gene expression.

Researchers found that the degree of methylation in the transposon's promoter is related to the amount of yellow pigment deposited in the hair shaft and that the amount of methylation varies from individual to individual. The result is variation in coat color phenotypes even in genetically identical mice (**Figure 19.13**). In these mice, coat colors range from yellow (unmethylated promoter) to pseudoagouti (highly methylated promoter). In addition to a gradation in coat color, there is also a gradation in body weight. Yellow mice are more obese than the brown pseudoagouti mice and are more likely to be diabetic. Alleles such as A^{vy} that show variable expression from individual to individual in genetically identical strains caused by different patterns of epigenetic modifications to the alleles are called *metastable epialleles*. "Metastable" refers to the variable nature of the epigenetic modifications, and "epiallele" refers to the heritability of the epigenetic status of the altered gene. In other words, the epigenetic modifications to the A^{vy} allele can be passed on to offspring; this is an example of transgenerational inheritance.

To evaluate the role of environmental factors in modifying the epigenome, the diet of pregnant A^{vy} mice was supplemented with methylation precursors, including folic acid, vitamin B_{12}, and choline. In the offspring, variation in coat color was reduced and shifted toward the pseudoagouti (highly methylated) phenotype. The shift in coat color was accompanied by increased methylation of the transposon's promoter. These findings have applications to epigenetic diseases in humans. For example, the risk of one form of colorectal cancer is linked directly to increased methylation of the DNA repair gene *MLH1*.

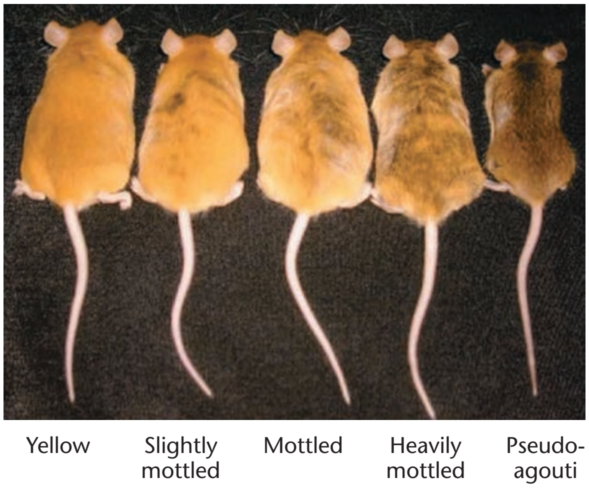


FIGURE 19.13 Variable expression of yellow phenotype in genetically identical mice caused by diet-related epigenetic changes in the A^{vy} allele.

Stress-Induced Behavior Is Heritable

A growing body of evidence shows that epigenetic changes, including alterations in DNA methylation and histone modification have important effects on behavioral phenotypes. In mice, two regions of the brain show preferential expression of maternal or paternal alleles. Upward of 1000 genes in the developing brain are imprinted, supporting the idea that epigenetic mechanisms operating in different regions of the brain may represent a major form of behavioral regulation.

In humans, epigenetic changes have been documented during the progression of neurodegenerative disorders and in neuropsychiatric diseases, both of which show altered behavioral phenotypes. Epigenetic changes to the nervous system occur in Alzheimer disease, Parkinson disease, Huntington disease, and in schizophrenia and bipolar disorder. However, because the phenotypes in these disorders are influenced by a number of factors including genetic predispositions, events in prenatal development, and prenatal and postnatal environmental effects, it is not yet possible to define a cause-and-effect relationship between epigenomic changes and the onset and intensity of neural disorders.

One of the most significant findings in the epigenetics of behavior is that stress-induced epigenetic changes that occur prenatally or early in life can influence behavior (and physical health) later in adult life and can potentially be transmitted to future generations. For example, a classic study showed that newborn rats raised with low levels of maternal nurturing (low-MN) did not adapt well to stress and anxiety-inducing situations in adulthood. In rats and humans, the hypothalamic region of the brain mediates stress reactions by controlling levels of glucocorticoid hormones via the action of cell-surface glucocorticoid receptors (GRs).

In rats exposed to high levels of maternal nurturing care early in life (high-MN), GR expression is increased and adults are stress adaptive. However, low-MN rats had reduced levels of GR transcription and were less able to adapt to stress. The relevant observation was that the differences in GR expression were associated with differences in histone acetylation and DNA methylation levels in the GR gene promoter. Low-MN rats had significantly higher levels of promoter methylation than high-MN rats (**Figure 19.14**).

Subsequent research showed that differences in DNA methylation are present in hundreds of genes across the genome, all of which show differential expression in low-MN and high-MN adults. Significantly, in low-MN adults, drugs that lower methylation levels reversed the effect of poor early-life nurturing and improved their stress responses. Later studies showed that these behavioral phenotypes can be transmitted across generations. Female rats raised by more nurturing mothers are more attentive to their own newborns, whereas those raised by less nurturing mothers are much less attentive and less nurturing to their offspring.

Similar epigenetic changes triggered by prenatal or early childhood environmental factors may alter later behavior in humans. For example, a history of child abuse increases the risk of suicide later in life. One study examined epigenomic differences in brain tissue from two classes of suicide victims and in others who died suddenly of unrelated causes. One group of suicide victims had experienced childhood abuse, and the other had no history of child abuse. Those who died suddenly of unrelated causes also had no history of child abuse. High levels of GR gene promoter methylation were found in suicide victims with a history of child abuse, but not in the other two groups. These results are consistent with those found in experimental animals and suggest that parental care, epigenomic variation, GR expression, and adult behavior are linked in both rats and humans. Further research may lead to the development of drugs to treat depression and help prevent suicide in humans.

Now Solve This

19.2 When considering transgenerational inheritance, it is important to consider whether the epigenetic marks controlling the phenotype are inherited through the gametes (sperm and eggs) or are laid down early in postnatal life by behavioral interaction with a parent. What experimental approaches would you use to determine which of these mechanisms operates in transmitting nurturing behavior in rats?

- **Hint:** *This problem asks you to design a set of experiments to analyze epigenetic marks in gametes and in early postnatal life. The key to the solution is to first consider whether the marks are in DNA or in histones and then to design experiments to test for the presence or absence of these epigenetic marks.*

19.5 Epigenome Projects and Databases

As the role of the epigenome in disease has become increasingly clear, researchers across the globe have formed multidisciplinary projects to map all the epigenetic changes that occur in the normal genome and established databases to study the role of the epigenome in specific diseases. We will discuss some of these projects and their goals and will summarize the major findings of a few of the large-scale projects.

The NIH Roadmap Epigenomics Project was established to elucidate the role of epigenetic mechanisms in

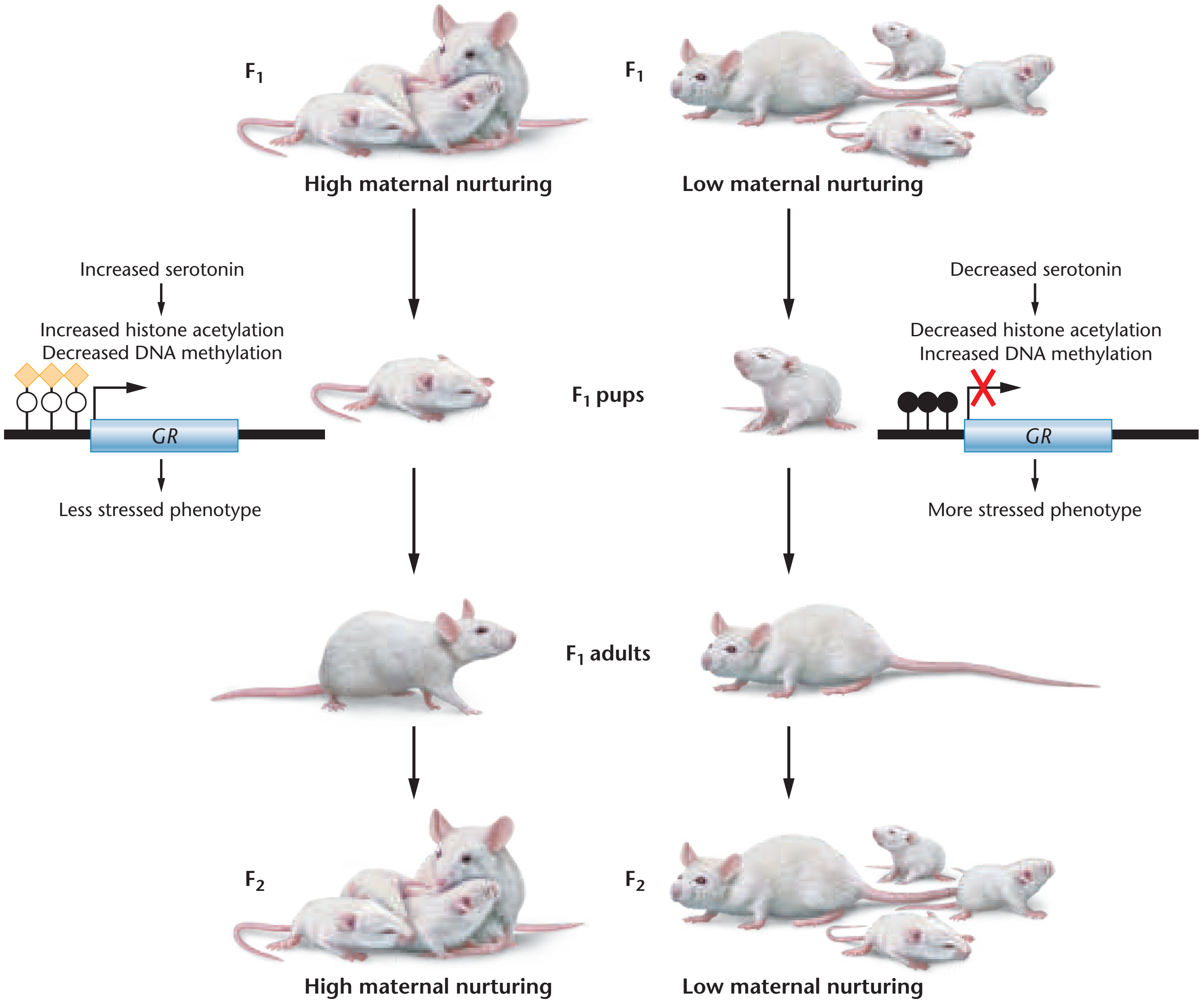


FIGURE 19.14 Style of maternal care is transmitted across rat generations through epigenetic events that take place early in postnatal life. High maternal nurturing induces high levels of serotonin in the brain, leading to DNA hypomethylation, histone acetylation, and increased expression of GR. In adulthood, high levels of GR expression increase adaptation to stress and, in females, passes on the high-MN phenotype. Rat pups experiencing low levels of maternal nurturing had higher levels of promoter methylation and reduced levels of GR expression. In adulthood, this led to poor stress adaptation and, in females, perpetuation of low levels of nurturing her pups.

human biology and disease. The project has two main goals: (1) provide a set of at least 1000 reference epigenomes in a range of cell types from healthy and diseased individuals, and (2) delineate the epigenetic differences in conditions such as Alzheimer disease, autism, and schizophrenia. The genome-wide data collected are classified into five categories: (1) histone modifications, (2) DNA methylation, (3) open and closed chromatin configurations, (4) expression levels of protein-coding genes, and (5) expression profiles for small nuclear RNAs, such as miRNA.

In 2015, the project published an analysis of the first 111 reference genomes collected, representing the most comprehensive map of the human epigenome to date. One of the important results of this study establishes that genetic variants associated with complex human disorders such as Alzheimer disease, cancer, and autoimmune disorders are enriched in tissue-specific epigenomic marks, identifying relevant cell types associated with these and other disorders. Studies on these cell types will open new areas of research into the molecular basis of complex traits. Other findings in this analysis contribute to studies of gene regulation, genome evolution, and genetic variation. The Human Epigenome Atlas, which is part of the Roadmap Project, collects and catalogs detailed information about epigenomic modifications at specific loci, in different cell types, different physiological states, and different genotypes. These data allow researchers to perform comparative analysis of epigenomic data across genomic regions or entire genomes.

The International Human Epigenome Consortium (IHEC) is a global program established to coordinate the collection of epigenome maps for 1000 human cell populations. Several projects are contributing to the program, each specializing in different cell types and/or approaches. The U.S. Reference Epigenome Mapping Centers are using stem cells and tissue samples from healthy donors, and the Germany-based DEEP Project is collecting 70 reference epigenomes of human and mouse tissues associated with metabolic and inflammatory diseases. In addition to epigenome mapping, the DEEP project is analyzing the contributions of specific cell types to disease. The European BLUEPRINT Project is collecting epigenomic profiles from several different types of blood cells related to specific diseases.

In addition, the IHEC is developing bioinformatics standards, as well as tools to organize, store, access, display, and analyze the epigenomic data gathered in this international effort. The data from the IHEC is available through a number of data portals including the GEO database of the U.S. National Center for Biotechnology Information, the European Bioinformatics Institute (EBI), and the Epigenome Atlas based at Baylor University.

To complement the efforts of IHEC in mapping the epigenomes of primary cell lines collected directly from tissues, the Encyclopedia of DNA Elements (ENCODE) project is focused on collecting epigenome maps for cell lines grown under laboratory conditions. To compare the epigenomes of normal cells with cancer cells, the International Cancer Genome Consortium (ICGC) is mapping the epigenomes and the transcriptome profiles of 50 different cancer types.

Although these projects are still in progress, the information already available strongly suggests that we are on the threshold of a new era in genetics, one in which we can study the development of disease at the genomic level and understand the impact of epigenetic factors on gene expression.

CASE STUDY Food for Thought

A couple well informed about the epigenetic effects of nutrition and dietary supplements such as choline and folate decided to take supplements before trying to get pregnant. Their first child was born with hair and eye colors very different from the parents and has been consistently praised by teachers for learning ability and memory. The couple attributes these characteristics to their intake of epigenetic supplements. Now they wish to have another child and have informed their physician that they intend to greatly increase the amount of supplements they take to increase the epigenetic effects in their child. The physician cautioned them that studies available in human trials show no effect of choline on learning and that excess intake of folate during pregnancy may be associated with an increased risk of autism spectrum disorder in the child. The couple is convinced that these studies are not conclusive, and while more research is needed, they intend to pursue their plan of increasing their intake of supplements.

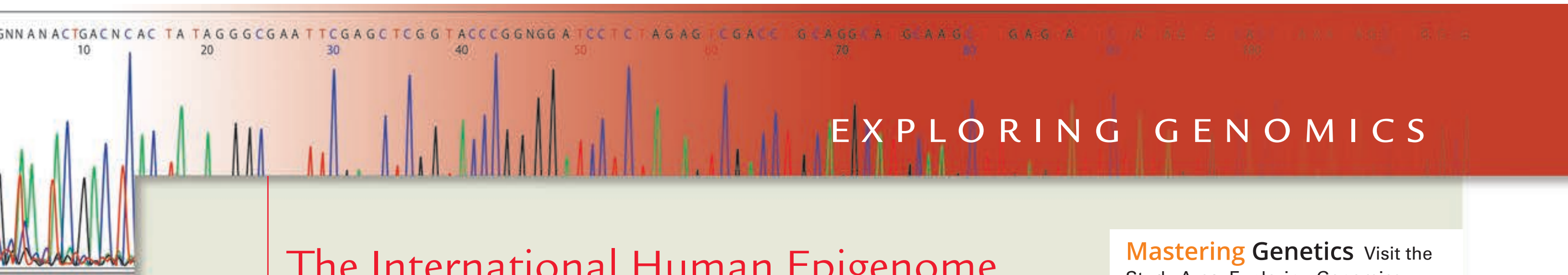
1. What would you say to the parents about the child's traits that they ascribe to epigenetics?
2. What tests might help establish whether the traits are genetic or epigenetic?
3. What ethical issues should the physician address with the parents about the use of supplements at high dose levels?

For related reading, see Geraghty, A., et al. (2015). Nutrition during pregnancy impacts offspring's epigenetic status—Evidence from human and animal studies. *Nutr. Metab. Insights* 8(S1):41–47.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. DNA methylated gene regions are highly condensed and genetically inactive. Levels of CpG methylation are high in promoter regions.
2. Active gene regions have an open or uncondensed chromatin structure, low levels of promoter and DNA methylation, and distinctive histone modifications (acetyl groups and methylation).
3. Patterns of DNA methylation in adult cells parallel cell function, chromatin structure, and gene activation.
4. Most DNA methylation patterning in the genome is removed at fertilization and reestablished during early embryogenesis.
5. Imprinted genes escape the pattern of demethylation and remethylation during early embryogenesis and show parent-specific patterns of methylation and activity of either the maternal or paternal allele.
6. X chromosome inactivation is an example of monoallelic expression and is correlated with lncRNA action and condensation of the chromosome.
7. Monoallelic expression of autosomal genes results in different expression profiles in cells within a given tissue, contributing to phenotypic variation.
8. Gene and CpG island methylation alterations are characteristic of cancer.
9. Specific epigenetic mechanisms active in cancer are the basis for the development of a new generation of chemotherapy drugs.
10. Environmental agents can alter the epigenome and lead to heritable altered patterns of gene expression.



EXPLORING GENOMICS

Mastering Genetics Visit the Study Area: Exploring Genomics

The International Human Epigenome Consortium (IHEC)

The International Human Epigenome Consortium (IHEC) was created in 2010 and involves member scientists from the United States, Canada, China, Germany, Japan, Korea, and the European Union. A primary goal of the IHEC is to coordinate development of reference maps for epigenomes of selected human cells and tissues in various stages of cellular development that will further an understanding of epigenetics and its impacts on human health and disease.

The IHEC also seeks to facilitate dissemination of epigenome data to other researchers around the world to help advance improvements in human health. The journal *Cell* created a portal to current key research papers that have emerged from the IHEC. Visit http://www.cell.com/consortium/IHEC for access to recent articles highlighting human epigenome studies for cancer, development, immunity, and many other topics.

Epigenome databases from IHEC members around the world, including the ENCODE project discussed in this chapter (and Chapter 21), contribute to the resources publically available through the Consortium. In this exercise, we will explore aspects of the **IHEC Data Portal**.

IHEC Data Portal

The IHEC Data Portal combines high-throughput sequencing data with extensive sets of epigenetic and transcriptome data sets from human tissue and animal disease models.

1. Access the IHEC homepage at http://ihec-epigenomes.org/. This site provides access to many different resources that may be of interest to you as you learn about epigenetics.

2. At the top of the IHEC homepage, click the “IHEC Data Portal” tab.

3. Once at the Portal, you will be on the “Overview” page and see a display of different IHEC databases that you can access through the portal. Notice these resources are clustered by consortium, tissue, or assay category.

4. From the “Genome” dropdown menu, select “Human (hg19).” (Note: hg19 and hg38 are two similar databases that display information in slightly different ways.) From the adjacent “Display” dropdown menu, select “Epigenomes.”

5. Using the “By Tissue” pie chart, click on “Other” (the color of this region of the pie chart will change to verify you selected it), and then click the “View Selected” box below the pie chart.

6. The next screen that displays is the Data Grid view. Rows on the left side of the grid display a listing of different human tissues analyzed. Columns of the grid indicate histone modifications (such as acetylation or methylation), genome methylation, and transcriptome data.

   There is a wealth of information available in these grids, so it is important that you simplify your data views and do not try to visualize too much information simultaneously.

7. Let’s explore an example of epigenome data available from a segment of the genome in human adipose tissue.

   The third tissue sample listed here is called “Adipose-Tissue-other.” For this tissue, click on the #1 box in the “H3K27ac” column. Then scroll down the grid until you see the “Visualize in Genome Browser” button and click this button. A new screen will appear taking you to the University of California, Santa Cruz Genome Browser site. Explore the data box that appears at the top of this site (most elements in the box can be clicked on or dragged for more information), just beneath the chromosome map, and answer the following questions:

   a. Which human chromosome is displayed? Are you viewing the p arm or q arm of this chromosome?

   b. Find the “RefSeq Genes” portion of this data box. You will see that this section of the chromosome includes the human *SOD* gene. What is this gene? What protein product does it encode, and what is the function of this protein? What inherited disorder is associated with mutations in the *SOD* gene?

   c. How many introns and exons of the *SOD* gene are represented here? Click on exon 1 for more information about the SOD protein.

   d. Scan the H3K27ac plot shown. What is the significance of H3K27ac marks in the genome? From these data, is H3K27ac more prevalent in the region of DNA encoding exon 1 and the transcription start site of the *SOD* gene or at other locations in the genome? Based on what you know about epigenetics, does this H3K27ac pattern make sense to you? Explain your answers.

   Notice that this view also shows:

   - Transcription (Txn) factor binding sites as determined experimentally by ChIP-seq analysis (recall that we discussed ChIP-seq in Chapter 21)
   - DNase I hypersensitivity regions of the genome determined experimentally
   - SNPs in this region of the chromosome

8. Now that you have been introduced to the IHEC Portal, spend some time exploring another portion of the epigenome for a tissue of interest to you. Be sure to reset the Data Grid before you begin a new search.

INSIGHTS AND SOLUTIONS

1. The epigenetic silencing of one of the homologous alleles of an autosomal gene is thought to be a random event. Given that upward of 20 percent of autosomal genes examined in several studies show monoallelic expression (MAE), what mechanisms might explain how monoallelic expression arises?

SOLUTION: Recall from Section 19.2 that four different specific patterns of monoallelic expression are possible: (1) both alleles are methylated and silent; (2) the paternal allele is methylated and silenced but the maternal allele is not methylated and is active; (3) the maternal allele is methylated and silenced but the paternal allele is not methylated and active; or (4) neither allele is methylated and both are active (biallelic expression).

Because the epigenetic mark H3K36me3 is associated with transcriptional activity and the H3K27me3 is associated with gene silencing, the presence of these marks is a reliable predictor of gene activity. This means that at some stage of development, one allele of a gene was randomly selected to be silenced by marking with the H3K27me3 marker, while the homologous allele received the H3K36me3 mark and is active. From this information, two general mechanisms present themselves: (1) If both alleles of a given gene are expressed (biallelic expression), then the silencing of one active allele would result in MAE. (2) If both alleles of a given gene are silent, activation of one allele would result in MAE. In either case, MAE might arise in a specific time or developmental window early in embryonic development when epigenetic marks are being laid down. This might be coupled with a feedback mechanism that inhibits the modification of the other allele, leaving one allele active and one silenced. While MAE is a widespread genomic phenomenon, presently, little is known about the process of selecting which allele to activate or silence, the timing of these modifications, or how they are maintained through rounds of mitosis.

2. Many cancers are associated with inactivation of tumor-suppressor genes. These changes can be of two kinds: genetic mutations that inactivate the suppressor gene, or inactivation by DNA methylation. The predominant model for tumor-suppressor gene methylation is the *stochastic model*, which supposes that methylation of the gene occurs by chance. The resulting cell escapes cell-cycle control and, having a selective growth advantage, eventually forms a malignant tumor. However, in another model, called the *instructional model*, an oncogene initiates a series of molecular events that recruits a transcription factor, which in turn, recruits a DNA methylase to bind to the tumor-suppressor gene, which is silenced by methylation.

Until recently, there has been little experimental evidence for the instructional model. However, in the last several years, examples of oncogene-directed silencing of tumor-suppressor genes have been accumulating. It now appears that instructive methylation of tumor-suppressor genes may be a general mechanism in cancer biology.

What are some potential therapeutic models for reversing methylation that derive from both the stochastic model and the instructional model of tumor-suppressor methylation?

SOLUTION: The stochastic model supposes that tumor-suppressor genes are randomly selected by an unknown mechanism for silencing, and once silenced, this condition is maintained in subsequent cell divisions by DNA methylase re-creating the methylation pattern of the copied DNA strand onto the newly synthesized DNA strand during the S phase.

To reactivate the tumor-suppressor gene in the stochastic model, there are two possible approaches: selectively reverse methylation of the tumor-suppressor gene in cancer cells, or prevent methylation of the gene on a newly replicated DNA strand. While methylation requires only the action of a single enzyme, reversing cytosine methylation is a complex process requiring the action of a family of enzymes (the TET family) and a number of steps. Attempting to reverse methylation of a specific tumor-suppressor gene as a therapeutic treatment would be difficult. It would require a way of targeting a single gene and initiating demethylation without setting off widespread demethylation of other genes, which might have the effect of making the tumor grow more aggressively.

During DNA replication, the enzyme DNMT1 is responsible for copying the methylation pattern from the parental DNA strand to the new strand. Inhibitors of DNA methyltransferases, especially nucleoside analogs, have proven somewhat effective in reactivating silenced tumor-suppressor genes by preventing their methylation during the S phase. These inhibitors are currently used as part of a therapeutic strategy in treating cancer.

Working from the instructional model would allow intervention before the tumor-suppressor gene is methylated. Oncogenes that silence tumor-suppressor genes do so through a series of intermediate steps. In one cell line, the *RAS* oncogene silences the *Fas* tumor-suppressor gene in a 16-step process, using 28 intermediate molecules. Using small molecules to interfere with any of these steps could prevent methylation and silencing.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. HOW DO WE KNOW? In this chapter, we focused on epigenetic modifications to the genome that regulate gene expression. Several mechanisms are involved, and epigenetic control of gene expression is important in development, cancer, and modulating the genomic response to environmental factors. From the explanations given in the chapter,

(a) How do we know how methylation of promoters silences gene expression?

(b) What is the evidence that epigenetic changes are involved in cancer?

(c) How does an environmental factor like stress generate a response that is transmitted from generation to generation?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 433. These concepts relate to the multiple ways epigenetic modifications of the genome lead to alterations in the pattern of gene expression. The fifth concept in the list describes how epigenetic changes, including DNA methylation and histone modifications, contribute to the causes and maintenance of cancer. Write a short essay describing how epigenetic changes in cancer cells contribute to the development and maintenance of cancers.

3. What are the major mechanisms of epigenetic genome modification?

4. What parts of the genome are reversibly methylated? How does this affect gene expression?

5. Identical twins each carry the same genome, but over time, can develop different phenotypes. How can you explain this?

6. What are the possible roles of proteins in histone modification?

7. Describe how reversible chemical changes to DNA and histones are linked to chromatin modification.

8. Why are changes in nucleosome spacing important in changing gene expression?

9. What are the similarities and differences in the two types of ncRNAs involved in epigenetic control of gene expression?

10. How do microRNAs regulate epigenetic mechanisms during development?

11. What are the functions of lncRNAs in epigenetic regulation? Describe each in detail.

12. What is the histone code?

13. What are the differences and similarities among the three classes of monoallelic gene expression?

14. What is the role of imprinting in human genetic disorders?

15. Imprinting disorders do not involve changes in DNA sequence, but only the methylated state of the DNA. Does it seem likely that imprinting disorders could be treated by controlling the maternal environment in some way, perhaps by dietary changes?

16. Should fertility clinics be required by law to disclose that some assisted reproductive technologies (ARTs) can result in epigenetic diseases? How would you and your partner balance the risks of ART with the desire to have a child?

17. How can the role of epigenetics in cancer be reconciled with the idea that cancer is caused by the accumulation of genetic mutations in tumor-suppressor genes and proto-oncogenes?

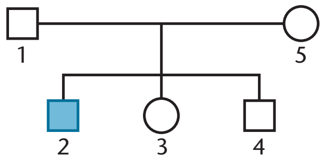
18. How are mutations in histone acetylation (HAT) genes linked to cancer?

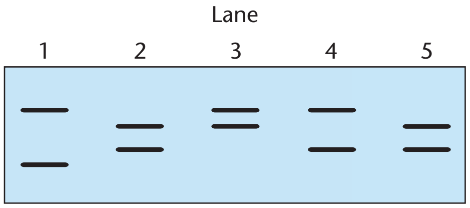
19. A developmental disorder in humans called spina bifida is a neural tube defect linked to a maternal diet low in folate during pregnancy.
(a) What does this suggest about the cause of spina bifida?
(b) Does this exclude genetic mutations as a cause of this condition?
(c) Should researchers be looking for mutant alleles of genes that control formation and differentiation of the neural tube?

20. Trace the relationship between the methylation status of the glucocorticoid receptor gene and the behavioral response to stress.

Extra-Spicy Problems

21. Prader–Willi syndrome (PWS) is a genetic disorder with a clinical profile of obesity, intellectual disability, and short stature. It can be caused in several ways. Most common is a deletion on the paternal copy of chromosome 15, but it can also be caused by an epigenetic imprinting disorder, and uniparental disomy, an event in which the affected child receives two copies of the maternal chromosome 15. A child with PWS comes to your clinic for a diagnosis of the molecular basis for this condition. The gel below shows the results of testing with short tandem repeats (STRs) from the region of chromosome 15 associated with the disorder.





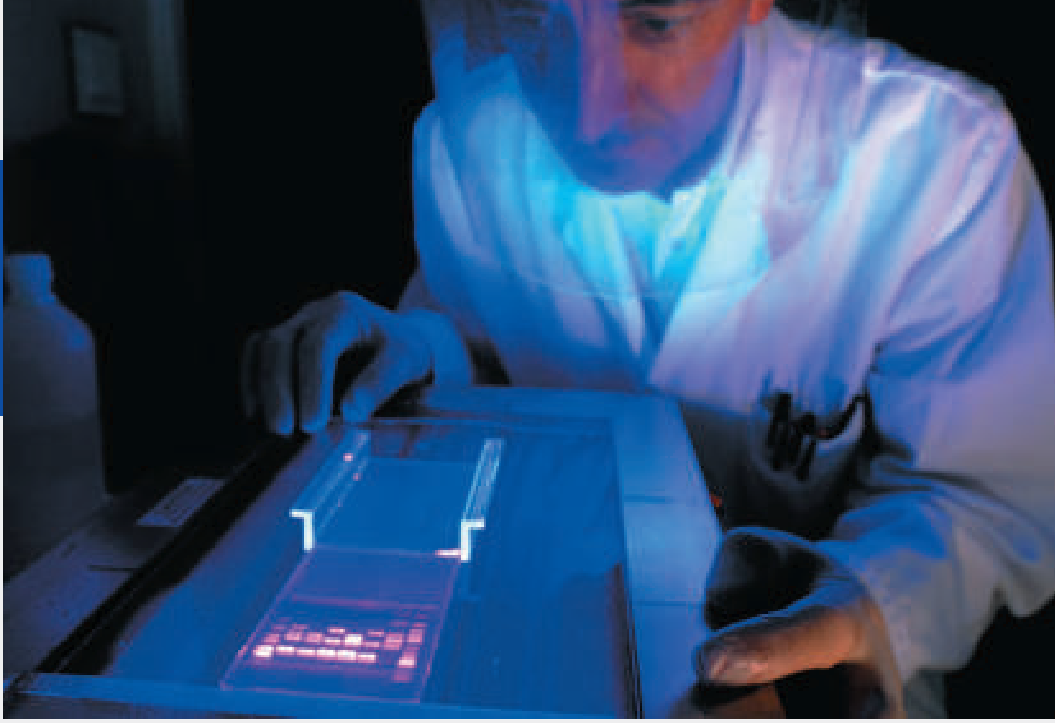
(a) Is this case caused by a deletion in the paternal copy of chromosome 15? Explain.
(b) Based on your interpretation of the data, what is the cause of PWS in this case? Explain your reasoning.

22. From the data in Table 19.3, draw up a list of histone H3 modifications associated with gene activation. Then draw up a list of H3 modifications associated with repression.
(a) Are there any overlaps on the lists?
(b) Are these overlaps explained by different modifications?
(c) If not, how can you reconcile these differences?

23. Amino acids are classified as positively charged, negatively charged, or electrically neutral.
(a) Which category includes lysine?
(b) How does this property of lysine allow it to interact with DNA?
(c) How does acetylation of lysine affect its interaction with DNA, and how is this related to the activation of gene expression?

24. Methylation of H3K9 by itself silences genes, but if H3K4 and H4K20 are also methylated, the combination of modifications stimulates transcription. What conclusions can you draw about this?

20



Recombinant DNA Technology

A researcher examines an agarose gel containing separated DNA fragments stained with the DNA-binding dye ethidium bromide and visualized under ultraviolet light.

CHAPTER CONCEPTS

- Recombinant DNA technology creates combinations of DNA sequences from different sources.
- A common application of recombinant DNA technology is to clone a DNA segment of interest.
- Specific DNA segments are inserted into vectors to create recombinant DNA molecules that are transferred into eukaryotic or bacterial host cells, where the recombinant DNA replicates as the host cells divide.
- DNA libraries are collections of cloned DNA and were historically used to isolate specific genes.
- DNA segments can be quickly amplified and cloned millions of times using the polymerase chain reaction (PCR).
- Nucleic acids can be analyzed using a range of molecular techniques.
- Sequencing reveals the nucleotide composition of DNA, and major improvements in sequencing technologies have rapidly advanced many areas of modern genetics research, particularly genomics.
- Gene knockout methods and transgenic animals have become invaluable for studying gene function *in vivo*.
- Gene editing methods demonstrate value in basic research and exciting potential for clinical applications of genetic technology.
- Recombinant DNA technology has revolutionized our ability to investigate the genomes of diverse species and has led to the modern revolution in genomics.

In 1971, a paper published by Kathleen Danna and Daniel Nathans marked the beginning of the recombinant DNA era. The paper described the isolation of an enzyme from bacteria and the use of the enzyme to cleave viral DNA at specific nucleotide sequences. It contained the first published photograph of DNA cut with such an enzyme, now called a restriction enzyme.

Using restriction enzymes and a number of other resources, researchers of the mid- to late 1970s developed various techniques to create, replicate, and analyze **recombinant DNA** molecules—DNA created by joining together pieces of DNA from different sources. These techniques, called **recombinant DNA technology** and often known as "gene splicing" in the early days, marked a major advance in research in molecular biology and genetics, allowing scientists to isolate and study specific DNA sequences. For their contributions to the development of this technology, Nathans, Hamilton Smith, and Werner Arber were awarded the 1978 Nobel Prize in Physiology or Medicine.

The power of recombinant DNA technology is astonishing, enabling geneticists to identify and isolate a single gene or DNA segment of interest from a genome. Through cloning, large quantities of identical copies of this specific DNA molecule can be produced. These identical copies, or **clones,** can then be manipulated for numerous purposes, including conducting research on the structure and organization of the DNA, studying gene expression, studying protein products to understand their structure and function, and producing important commercial products from the protein encoded by a gene.

The fundamental techniques involved in recombinant DNA technology subsequently led to the field of genomics, enabling scientists to sequence and analyze entire genomes. Note that some of the topics discussed in this chapter are explored in greater depth later in the text (see Special Topic Chapters 2—DNA Forensics, 4—Genetically Modified Foods, and 5—Gene Therapy). In this chapter, we survey basic methods of recombinant DNA technology used to isolate, replicate, and analyze DNA.

20.1 Recombinant DNA Technology Began with Two Key Tools: Restriction Enzymes and Cloning Vectors

Although natural genetic processes such as crossing over produce recombined DNA molecules, the term *recombinant DNA* is reserved for molecules produced by artificially joining DNA obtained from different sources. We begin our discussion of recombinant DNA technology by considering two important tools used to construct and amplify recombinant DNA molecules: DNA-cutting enzymes called **restriction enzymes** and **cloning vectors.** The use of restriction enzymes and cloning vectors was largely responsible for advancing the field of molecular biology because a wide range of laboratory techniques are based on recombinant DNA technology.

Restriction Enzymes Cut DNA at Specific Recognition Sequences

Restriction enzymes are produced by bacteria as a defense mechanism against infection by bacteriophage. They *restrict* or prevent viral infection by degrading the DNA of invading viruses. More than 4300 restriction enzymes have been identified, and over 600 are commercially produced and available for use by researchers. A restriction enzyme recognizes and binds to DNA at a specific nucleotide sequence called a **recognition sequence** or **restriction site** (Figure 20.1). The enzyme then cuts both strands of the DNA within that sequence by cleaving the phosphodiester backbone. Scientists commonly refer to this as “digestion” of DNA. The usefulness of restriction enzymes is their ability to accurately and reproducibly cut DNA into fragments. Restriction enzymes represent sophisticated molecular scissors for cutting DNA into fragments of desired sizes.

Restriction sites are distributed randomly in the genome. The number of DNA restriction fragments produced by digesting DNA with a particular enzyme can be estimated from the number of times a given restriction enzyme cuts the DNA. Enzymes with a four-base recognition sequence—such as the enzyme *Alu*I, which recognizes the sequence AGCT—will cut, on average, every 256 base pairs ($4^n = 4^4 = 256$) if all four nucleotides are present in equal proportions, producing many small fragments. The actual fragment sizes produced by DNA digestion with a given restriction enzyme vary because of variability in the number of recognition sequences in relation to one another.

Recognition sequences exhibit a form of symmetry described as a **palindrome:** The nucleotide sequence reads

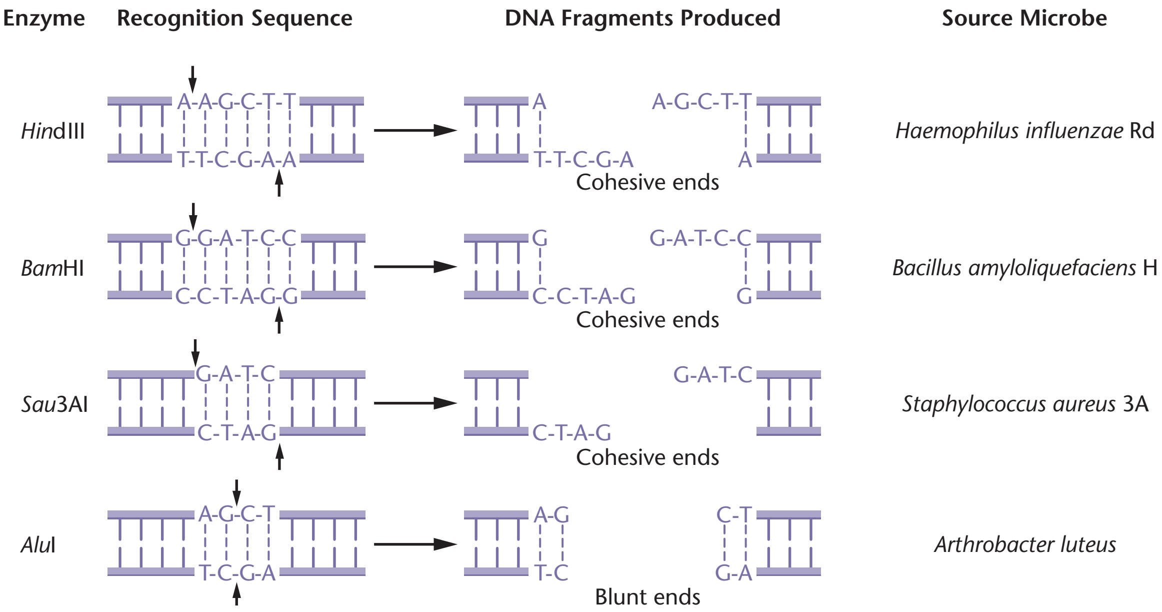


FIGURE 20.1 Common restriction enzymes, with their recognition sequence, DNA cutting patterns, and source microbes. Arrows indicate the location in the DNA cut by each enzyme.

the same on both strands of the DNA when read in the 5′ to 3′ direction. Each restriction enzyme recognizes its particular recognition sequence and cuts the DNA in a characteristic cleavage pattern (see Figure 20.1). The most common recognition sequences are four or six nucleotides long, but some contain eight or more nucleotides. Enzymes such as *Hind*III make offset cuts in the DNA strands, thus producing fragments with single-stranded overhanging ends called *cohesive ends* (or *"sticky" ends*), while others such as *Alu*I cut both strands at the same nucleotide pair, producing DNA fragments with double-stranded ends called *blunt-end* fragments.

One of the first restriction enzymes to be identified was isolated from *Escherichia coli* strain R and was designated *Eco*RI. DNA fragments produced by *Eco*RI digestion have cohesive ends because they can base-pair with complementary single-stranded ends on other DNA fragments cut using *Eco*RI. When mixed together, single-stranded ends of DNA fragments from different sources cut with the same restriction enzyme can **anneal,** or stick together, by hydrogen bonding of complementary base pairs in single-stranded ends (**Figure 20.2**). Addition of the enzyme **DNA ligase**—recall the role of DNA ligase in DNA replication as discussed earlier in the text (see Chapter 11)—to DNA fragments will seal the phosphodiester backbone of DNA to covalently join the fragments together to form recombinant DNA molecules.

Scientists often use restriction enzymes that create cohesive ends since the overhanging ends make it easier to combine fragments. Blunt-end ligation is more technically challenging because it is not facilitated by hydrogen bonding, but a scientist can ligate fragments digested at different sequences by different blunt-end generating enzymes.

DNA Vectors Accept and Replicate DNA Molecules to Be Cloned

Scientists recognized that DNA fragments resulting from restriction-enzyme digestion could be copied or cloned if they also had a technique for replicating the fragments. Thus, a second key tool that allowed DNA cloning was the development of **cloning vectors,** DNA molecules that accept DNA fragments and replicate these fragments when vectors are introduced into host cells.

Many different vectors are available for cloning. Vectors differ in terms of the host cells they are able to enter and in the size of DNA fragment inserts they can carry, but most DNA vectors share several key properties.

- A vector contains several restriction sites that allow insertion of the DNA fragments to be cloned.
- Vectors must be capable of replicating in host cells independent of the host cell chromosome(s).

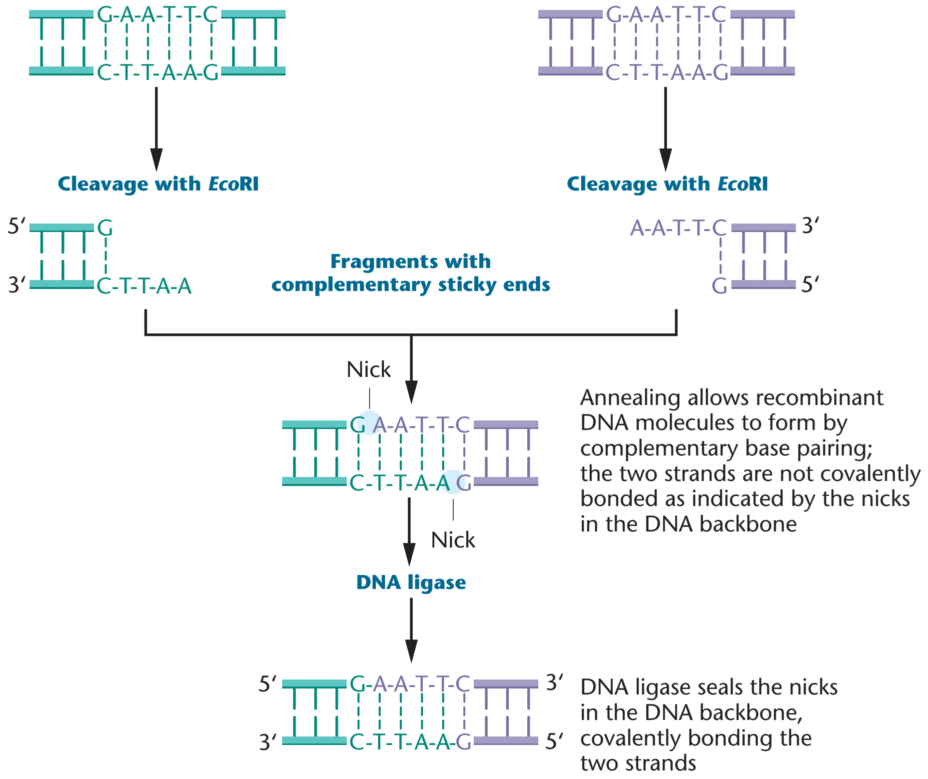

G-A-A-T-T-C
C-T-T-A-A-G
G-A-A-T-T-C
C-T-T-A-A-G
Cleavage with *Eco*RI
Cleavage with *Eco*RI
5′ G
3′ C-T-T-A-A
Fragments with complementary sticky ends
A-A-T-T-C 3′
G 5′
Nick
G A-A-T-T-C
C-T-T-A-A G
Nick
Annealing allows recombinant DNA molecules to form by complementary base pairing; the two strands are not covalently bonded as indicated by the nicks in the DNA backbone
DNA ligase
5′ G-A-A-T-T-C 3′
3′ C-T-T-A-A-G 5′
DNA ligase seals the nicks in the DNA backbone, covalently bonding the two strands

FIGURE 20.2 DNA from different sources is cleaved with *Eco*RI and mixed to allow annealing. The enzyme DNA ligase forms phosphodiester bonds between these fragments to create a recombinant DNA molecule.

- To make it possible to distinguish host cells that have taken up vectors from host cells that have not, the vector should carry a **selectable marker gene** (often an antibiotic resistance gene) or a **reporter gene** that encodes a protein which produces a visible effect, such as color or fluorescent light).
- Most vectors incorporate specific sequences that allow for sequencing inserted DNA.

Bacterial Plasmid Vectors

Genetically modified bacterial **plasmids** were the first vectors developed, and they are still widely used for cloning. Plasmid cloning vectors were derived from naturally occurring plasmids. Recall from earlier discussions (Chapter 6) that plasmids are extrachromosomal, double-stranded, circular DNA molecules that replicate independently from the chromosomes within bacterial cells [Figure 20.3(a)]. Plasmids have been extensively modified by genetic engineering to serve as cloning vectors. Many commercially prepared plasmids are readily available with a range of useful features [Figure 20.3(b)].

Plasmids are introduced into bacteria by the process of **transformation** (see Chapter 6). Two main techniques are widely used for bacterial transformation. One approach involves treating cells with calcium ions and using a brief heat shock to introduce the plasmid DNA into cells. The other technique, called *electroporation*, uses a brief, but high-intensity, pulse of electricity to move plasmid DNA into bacterial cells.

Only one or a few plasmids generally enter a bacterial host cell by transformation. But because plasmids have an *origin of replication (ori)* site that allows for plasmid replication, it is possible to produce several hundred copies of a plasmid in a single host cell. This greatly enhances the number of DNA clones that can be produced. Plasmid vectors have also been genetically engineered to contain a number of restriction sites for commonly used restriction enzymes in a region called the *multiple cloning site*. Multiple cloning sites allow scientists to clone a range of different fragments generated by many commonly used restriction enzymes.

Cloning DNA with a plasmid generally begins by cutting both the plasmid DNA and the DNA to be cloned with the same restriction enzyme (Figure 20.4). Typically, the plasmid is cut once within the multiple cloning site, converting the circular molecule into a linear vector. DNA restriction fragments from the DNA to be cloned are added to the linearized vector in the presence of DNA ligase. Sticky ends of DNA fragments anneal, joining the DNA to be cloned and the plasmid. DNA ligase is then used to create phosphodiester bonds to seal nicks in the DNA backbone, thus producing recombinant DNA, which is then introduced into bacterial host cells by transformation. Once inside the cell, plasmids replicate quickly to produce multiple copies.

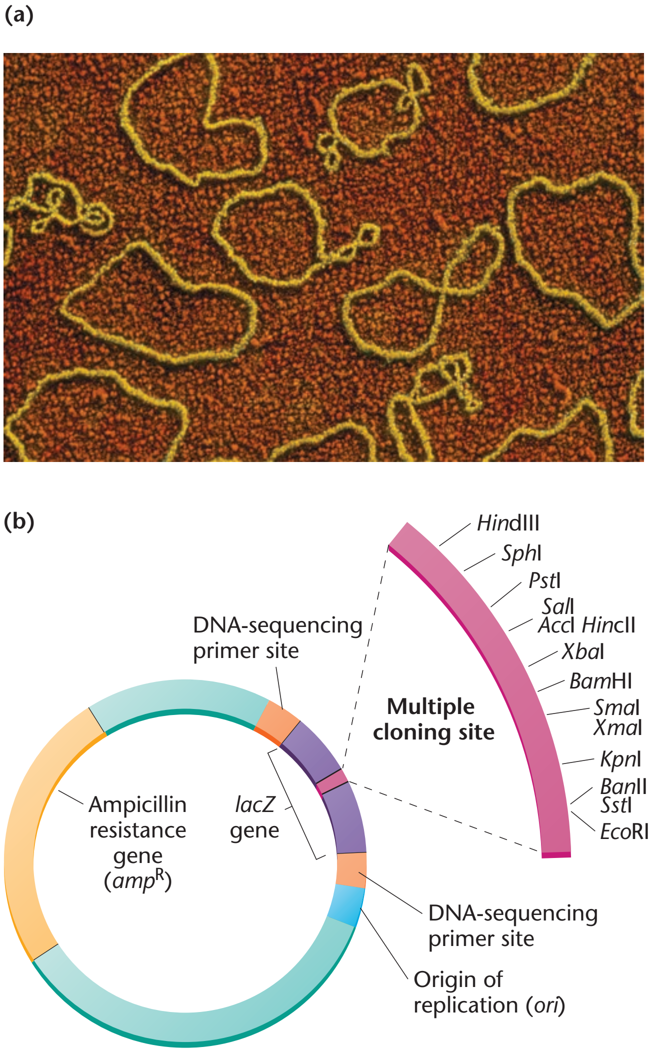

(a)
(b)
HindIII
SphI
PstI
SalI
AccI HincII
XbaI
BamHI
SmaI
XmaI
KpnI
BanII
SstI
EcoRI
DNA-sequencing primer site
Multiple cloning site
Ampicillin resistance gene (ampR)
lacZ gene
DNA-sequencing primer site
Origin of replication (ori)

FIGURE 20.3 (a) Color-enhanced electron micrograph of plasmids isolated from *E. coli.* (b) Diagram of a typical DNA cloning plasmid.

However, when cloning DNA using plasmids, not all plasmids will incorporate DNA to be cloned. For example, a plasmid cut with a restriction enzyme generating sticky ends can close back on itself (self-ligate) if cut ends of the plasmid rejoin. Obviously, such nonrecombinant plasmids are not desired. Also, during transformation, not all host cells will take up plasmids. Therefore, it is important that bacterial cells containing recombinant DNA can be readily identified in a cloning experiment. One way this is accomplished is through the use of selectable marker genes. Genes that provide resistance to antibiotics such as *amp*^{R} for ampicillin resistance and genes such as the *lacZ* gene are very effective selectable marker genes. Figure 20.5 provides an example of how the latter can be used to identify

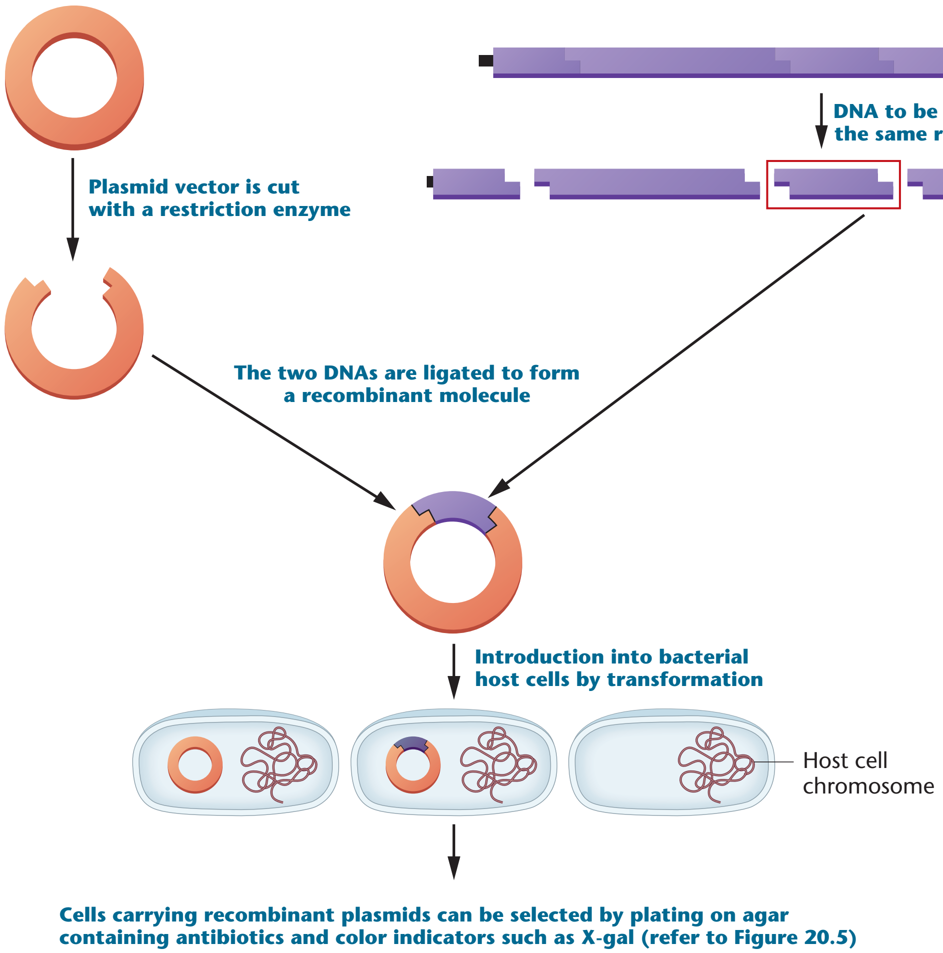

DNA to be cloned is cut with the same restriction enzyme
Plasmid vector is cut with a restriction enzyme
The two DNAs are ligated to form a recombinant molecule
Introduction into bacterial host cells by transformation
Host cell chromosome
Cells carrying recombinant plasmids can be selected by plating on agar containing antibiotics and color indicators such as X-gal (refer to Figure 20.5)

FIGURE 20.4 Cloning with a plasmid vector.

bacteria containing recombinant plasmids. This process is often referred to as **"blue-white" screening** for a reason that will soon become obvious.

In blue-white screening, a plasmid is used that contains the *lacZ* gene into which a multiple cloning site has been incorporated. The *lacZ* gene encodes the enzyme β-galactosidase, which, as you learned earlier in the text (see Chapter 16), is used to cleave the disaccharide lactose into its component monosaccharides glucose and galactose. Blue-white screening takes advantage of this enzymatic activity. Using this approach, one can easily identify transformed cells containing recombinant or nonrecombinant plasmids. If a DNA fragment is inserted anywhere in the multiple cloning site, the *lacZ* gene is disrupted and will not produce functional copies of β-galactosidase. The agar plates used in the assay contain an antibiotic—ampicillin in this case. Nontransformed bacteria cannot grow well on these plates because they do not have the *amp*^{R} gene and so the ampicillin kills these cells.

These agar plates also contain a substance called X-gal (technically 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside). X-gal is similar to lactose in structure. It is a substrate for β-galactosidase, and when it is cleaved by β-galactosidase it turns blue. As a result, bacterial cells carrying nonrecombinant plasmids (those that have self-ligated and thus do not contain inserted DNA) have a functional *lacZ* gene and produce β-galactosidase, which cleaves X-gal in the medium, and these cells turn blue. However, recombinant bacteria with plasmids containing an inserted DNA fragment will form white colonies when they grow on X-gal medium because the plasmids in these cells are not producing functional β-galactosidase (Figure 20.5, bottom). Bacteria in a colony are clones of each other—genetically identical cells with copies of recombinant plasmids. White colonies can be transferred to flasks of bacterial culture broth and grown in large quantities, after which it is relatively easy to isolate and purify recombinant plasmids from these cells.

Plasmids are still the workhorses for many applications of recombinant DNA technology, but they have a major limitation: Because they are small, they can only accept inserted pieces of DNA up to about 25 kilobases (kb) in size, and most plasmids can often only accept substantially smaller pieces. Therefore, as recombinant DNA technology developed and it became desirable to clone large pieces of DNA, other vectors were developed primarily for their ability to accept larger pieces of DNA and because they could be used with other types of host cells beside bacteria.

Other Types of Cloning Vectors

Phage vectors were among the earliest vectors used in addition to plasmids. These included genetically modified strains of bacteriophage λ—a double-stranded DNA virus. The genome of λ phage has been sequenced, and it has been modified to incorporate many of the important features of cloning vectors described earlier in this chapter, including a multiple cloning site. Phage vectors were popular for quite some time and are still in use today because they can carry inserts up to 45 kb—more than twice as long as DNA inserts in most plasmid vectors. DNA fragments are ligated into the phage vector to produce recombinant λ vectors that are subsequently packaged into phage protein heads *in vitro*, and then the phage are used to infect bacterial host cells growing on petri plates. Inside the bacteria, the vectors replicate and form many copies of infective phage, each of which carries a DNA insert. As they reproduce, they lyse their bacterial host cells, forming the clear spots known as plaques on the bacterial lawn (described in Chapter 6), from which phage can be isolated and the cloned DNA can be recovered.

Bacterial artificial chromosomes (BACs) and **yeast artificial chromosomes (YACs)** are two other examples of vectors that can be used to clone large fragments of DNA. For example, the mapping and analysis of large eukaryotic genomes, including the human genome, require cloning vectors that can carry very large DNA fragments such as segments of an entire chromosome. BACs are essentially very large but low copy number (typically one or two copies per bacterial cell) plasmids that can accept DNA inserts in the 100- to 300-kb range. Like natural chromosomes, a YAC has telomeres at each end, origins of replication, and a centromere. These components are joined to selectable marker genes and to a cluster of restriction-enzyme recognition sequences for insertion of foreign DNA. Yeast chromosomes range in size from 230 kb to over 1900 kb, making it possible to clone DNA inserts from 100 to 1000 kb in YACs. The

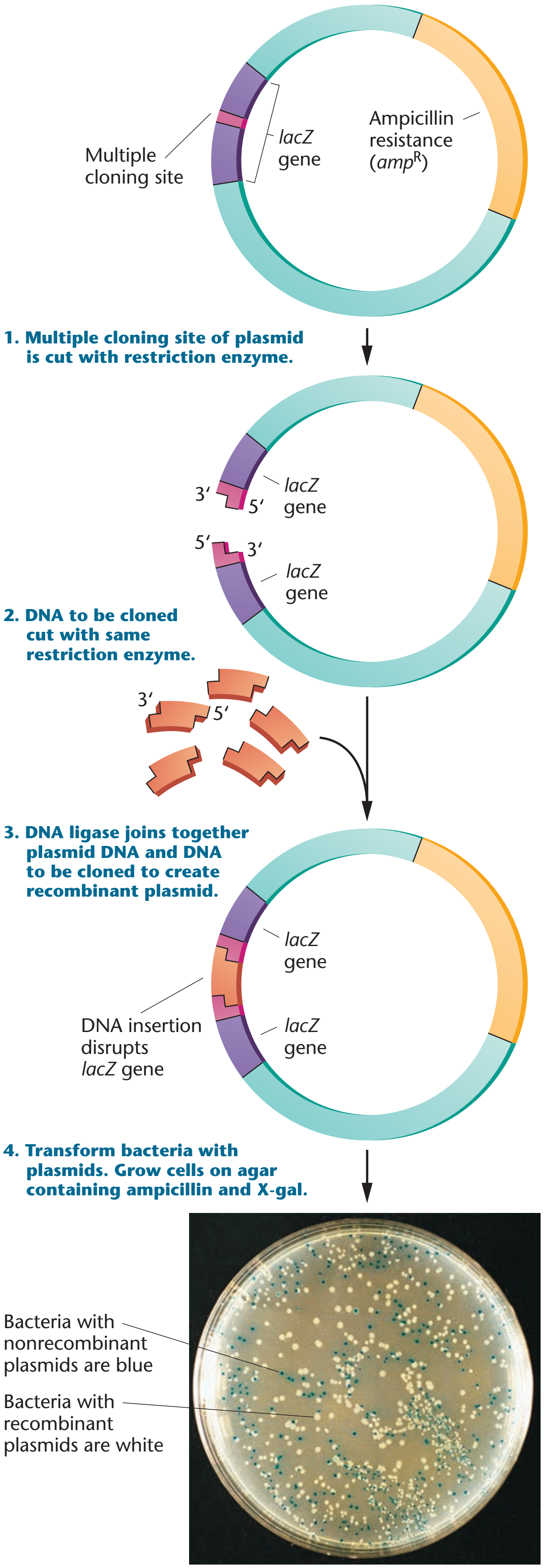

Multiple
cloning site
lacZ
gene
Ampicillin
resistance
(*amp*^{R})
1. Multiple cloning site of plasmid is cut with restriction enzyme.
3′
5′
lacZ
gene
5′
3′
lacZ
gene
2. DNA to be cloned cut with same restriction enzyme.
3′
5′
3. DNA ligase joins together plasmid DNA and DNA to be cloned to create recombinant plasmid.
lacZ
gene
DNA insertion
disrupts
lacZ gene
lacZ
gene
4. Transform bacteria with plasmids. Grow cells on agar containing ampicillin and X-gal.
Bacteria with
nonrecombinant
plasmids are blue
Bacteria with
recombinant
plasmids are white

FIGURE 20.5 In blue-white screening, DNA inserted into the multiple cloning site of a plasmid disrupts the *lacZ* gene so that bacteria containing recombinant DNA are unable to metabolize X-gal, resulting in white colonies that allow direct identification of bacterial colonies carrying DNA inserts to be cloned. (Bottom) Photo of a Petri dish showing the growth of bacterial cells after uptake of plasmids. Cells in blue colonies contain vectors without DNA inserts (nonrecombinant plasmids), whereas cells in white colonies contain vectors carrying DNA inserts (recombinant plasmids). Nontransformed cells did not grow into colonies due to the presence of ampicillin in the plating medium.

ability to clone large pieces of DNA in these vectors made them an important tool in the Human Genome Project (see Chapter 21).

Unlike the vectors described so far, **expression vectors** are designed to ensure mRNA expression of a cloned gene with the purpose of producing many copies of the gene's encoded protein in a host cell. This is an important distinction since most plasmids, phage vectors, and YACs only carry DNA and do not signal the cell to transcribe it into mRNA. Expression vectors are available for both bacterial and eukaryotic host cells and contain the appropriate sequences to initiate both transcription and translation of the cloned gene. For many research applications that involve studies of protein structure and function, producing a recombinant protein in bacteria (or other host cells) and purifying the protein is a routine approach, although it is not always easy to properly express a protein that maintains its biological function. The biotechnology industry also relies heavily on expression vectors to produce commercially valuable protein products from cloned genes, a topic we will discuss later in the text (see Chapter 22).

Introducing genes into plants is a common application that can be done in many ways, and we will discuss aspects of genetic engineering of food plants later in the text (see Special Topic Chapter 4—Genetically Modified Foods). One widely used approach to insert genes into plant cells involves a species of soil bacterium and a type of plasmid called a **Ti plasmid.**

Host Cells for Cloning Vectors

Besides deciding which DNA cloning vector to use, another cloning consideration worthy of discussion is which host cells are used to accept recombinant DNA for cloning. There are many different reasons why particular host cells are chosen for a recombinant DNA experiment, depending on the purpose of the work.

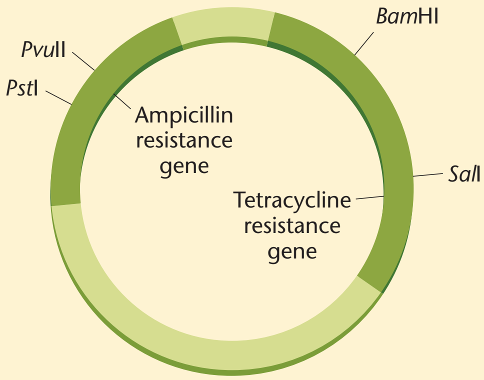
Whereas *E. coli* is widely used as a bacterial host cell of choice when working with plasmids to manipulate short fragments of DNA, the yeast *Saccharomyces cerevisiae* is extensively used as a host cell for the cloning and expression of eukaryotic genes. There are several reasons for choosing yeast:

1. Although yeast is a eukaryotic organism, it can be grown and manipulated in much the same way as bacterial cells.
2. The genetics of yeast has been intensively studied.
3. The genomes for popular strains of yeast have been sequenced.
4. To study the function of some eukaryotic proteins, it is necessary to use a host cell that can modify the protein, by adding carbohydrates, for example, after it has been synthesized, to convert it to a functional form (bacteria cannot carry out most of these modifications).
5. Yeast has been used for centuries in the baking and brewing industries, and is considered to be a safe organism for producing proteins for vaccines and therapeutic agents.

Choices for eukaryotic host cells also extend to a number of other cell types, including insect cells and human cells. A variety of different human cell types can be grown in culture and used to express genes and proteins. Such cell lines can also then be subjected to various approaches for gene or protein functional analysis, including drug testing for effectiveness at blocking or influencing a particular recombinant protein being expressed, particularly if the cell lines are of a human disease condition such as cancer.

NOW SOLVE THIS

20.1 A plasmid that is both ampicillin and tetracycline resistant is cleaved with *Pst*I, which cleaves within the ampicillin resistance gene. The cut plasmid is ligated with *Pst*I-digested *Drosophila* DNA to prepare a genomic library, and the mixture is used to transform *E. coli* K12.



(a) Which antibiotic should be added to the medium to select cells that have incorporated a plasmid?

(b) If recombinant cells were plated on medium containing ampicillin or tetracycline and medium with both antibiotics, on which plates would you expect to see growth of bacteria containing plasmids with *Drosophila* DNA inserts?

(c) How can you explain the presence of colonies that are resistant to both antibiotics?

HINT: *This problem involves an understanding of antibiotic selectable marker genes in plasmids and antibiotic DNA selection for identifying bacteria transformed with recombinant plasmid DNA. The key to its solution is to recognize that inserting foreign DNA into the plasmid vector disrupts one of the antibiotic resistance genes in the plasmid.*

20.2 DNA Libraries Are Collections of Cloned Sequences

Only relatively small DNA segments—representing just a single gene or even a portion of a gene—are produced by cloning DNA into smaller vectors, particularly plasmids. Even when several hundred genes are introduced into larger vectors such as BACs or YACs, one still needs a method for identifying the DNA pieces that were cloned. Consider this: In our cloning discussions so far, we have described how DNA can be inserted into vectors and cloned—a relatively straightforward process—but we have not discussed how a researcher knows what particular DNA sequence they have cloned. Simply cutting DNA and inserting it into vectors does not tell you what gene or sequences are being copied.

During the first several decades of DNA cloning, scientists created **DNA libraries,** which represent collections of cloned DNA. Depending on how a library is constructed, it may contain genes and noncoding regions of DNA. Generally speaking, there are two main types of libraries, genomic DNA libraries and complementary DNA (cDNA) libraries.

Genomic Libraries

Ideally, a **genomic library** consists of many overlapping fragments of the genome, with at least one copy of every DNA sequence in an organism's chromosomes, which in summary span the entire genome.

In making a genomic library, chromosomal DNA is extracted from cells or tissues and cut randomly with restriction enzymes, and the resulting fragments are inserted into vectors using techniques that we discussed in Section 20.1. Vectors in the genomic library may contain more than one gene or only a portion of a gene. Also, libraries built from eukaryotic cells will contain coding and noncoding segments of DNA such as introns.

Since some vectors (such as plasmids) can carry only a few thousand base pairs of inserted DNA, selecting the vector so that the library contains the whole genome in the smallest number of clones is an important consideration. This consideration becomes primary when working with large genomes such as the human genome. As a result, BACs and YACs were commonly used to accommodate the large sizes of DNA necessary to span the approximately 3 billion bp of human DNA (as in the Human Genome Project). If, for example, a human genome library was constructed using plasmid vectors with an average insert size of 5 kb, then more than 2.4 million clones would be required for a 99 percent probability of recovering any given sequence from the genome. Because of its size, this library would be difficult to use efficiently. However, if the library was constructed in, for example, a YAC vector with an average insert size of 1 Mb, then the library would only need to contain about 14,000 YACs, making it relatively easy to use.

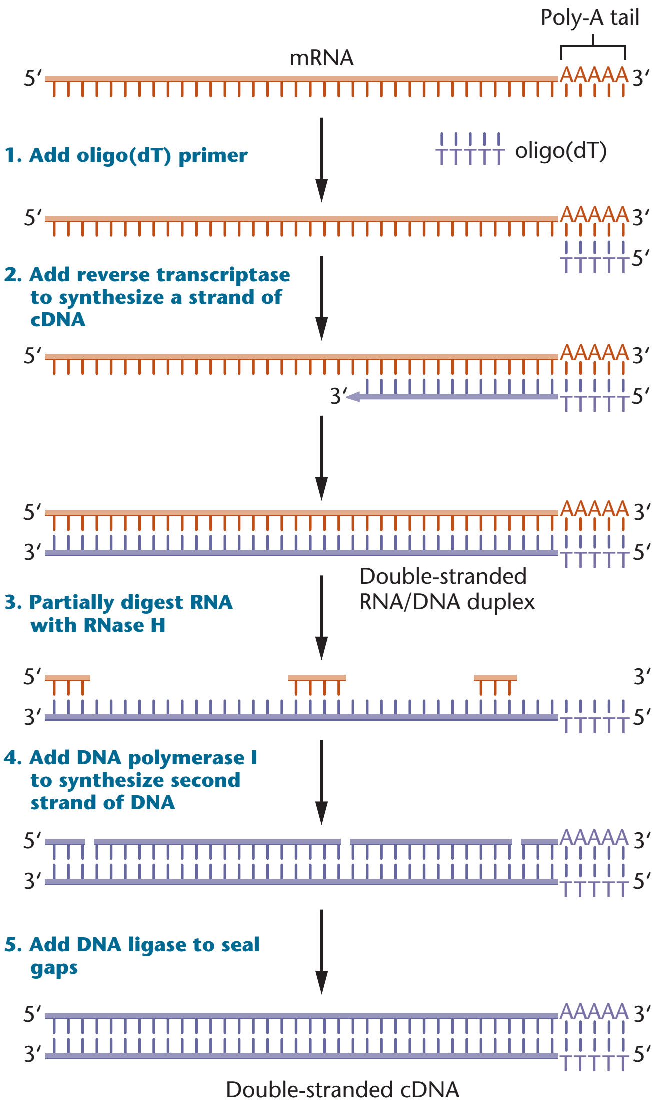
As you will learn later in the text (see Chapter 21), **whole-genome sequencing** approaches (see Figure 21.1) and new sequencing methodologies are replacing traditional genomic DNA libraries because they effectively allow one to sequence an entire genomic DNA sample without the need for inserting DNA fragments into vectors and cloning them in host cells. But the concept of a DNA library is still important for a number of modern applications. We will also consider later (Chapter 21) how DNA sequence analysis using bioinformatics allows one to identify protein-coding and noncoding sequences in cloned DNA.

Complementary DNA (cDNA) Libraries

Complementary DNA (cDNA) libraries offer certain advantages over genomic libraries and continue to be a useful methodology for specific approaches to gene cloning and other applications. This is primarily because a cDNA library contains DNA copies—called cDNA—which are made from mRNA molecules isolated from cultured cells or a tissue sample. A cDNA library therefore represents only the genes being expressed in cells at the time the library was made—unlike a genomic library, which contains all of the DNA, coding and noncoding, in a genome. This is a key point: cDNA libraries provide a snapshot, or catalog, of just the genes that were transcriptionally active in a tissue at a particular time.

As a result, cDNA libraries have been particularly useful for identifying and studying genes expressed in certain cells or tissues under certain conditions: for example, during development, cell death, cancer, and other biological processes. One can also use these libraries to compare expressed genes from normal tissues and diseased tissues. For instance, this approach has been widely used to identify genes involved in cancer formation, such as those genes that contribute to progression from a normal cell to a cancer cell and genes involved in cancer cell metastasis (spreading).

The initial steps required to prepare a cDNA library are shown in Figure 20.6. Key to the technique is the process of **reverse transcription.** Because most eukaryotic mRNAs have a poly-A tail at the 3′ end, a short oligo(dT) molecule is annealed to this tail to serve as a primer for initiating DNA synthesis by the enzyme reverse transcriptase. Reverse transcriptase uses the mRNA as a template to synthesize a complementary DNA strand (cDNA) and forms a double-stranded mRNA/cDNA duplex. The mRNA is then partially digested with the enzyme RNAse H to produce gaps in the RNA strand. The 3′ ends of the remaining mRNA serve as primers for DNA polymerase I, which synthesizes a second DNA strand. The result is a double-stranded cDNA



Poly-A tail
mRNA
5′
AAAAA 3′
1. Add oligo(dT) primer
oligo(dT)
5′
AAAAA 3′
5′
2. Add reverse transcriptase to synthesize a strand of cDNA
5′
AAAAA 3′
3′
5′
5′
AAAAA 3′
3′
5′
Double-stranded RNA/DNA duplex
3. Partially digest RNA with RNase H
5′
3′
3′
5′
4. Add DNA polymerase I to synthesize second strand of DNA
5′
AAAAA 3′
3′
5′
5. Add DNA ligase to seal gaps
5′
AAAAA 3′
3′
5′
Double-stranded cDNA

FIGURE 20.6 Producing cDNA from mRNA.

molecule that can be cloned into suitable vectors, usually plasmids.

Because one typically wouldn't know what restriction enzymes could be used to cut cDNA produced by the method just described, one usually needs to attach linker sequences to the ends of the cDNA in order to insert it into a plasmid. Linkers are short double-stranded oligonucleotides containing a restriction-enzyme recognition sequence (e.g., for *Eco*RI). After attachment to the cDNAs, the linkers are cut with *Eco*RI and ligated to vectors treated with the same enzyme.

Specific Genes Can Be Recovered from a Library by Screening

Genomic and cDNA libraries often consist of several hundred thousand different DNA clones, much like a large book library may have many books but only a few of interest to your studies in genetics. So how can libraries be used to locate a specific gene of interest? To find a specific gene, we need to identify and isolate only the clone or clones containing that gene. We must also determine whether a given clone contains all or only part of the gene we are trying to study.

For several decades an approach called *library screening* was routinely used to sort through a library and isolate specific genes of interest. Many of the first genes to be cloned and sequenced were identified this way. Library screening usually involves use of a **probe,** any DNA or RNA sequence that is complementary to some part of the target gene or sequence to be identified in a library. The probe will bind (hybridize) to any complementary DNA sequences present in one or more clones.

Probes are derived from a variety of sources—often related genes isolated from another species can be used if enough of the DNA sequence is conserved. For example, genes from rats, mice, or even *Drosophila* that have conserved sequence similarity to human genes can be used as probes to identify human genes during library screening.

A probe must be labeled or tagged in some way so that it can be identified. Initially probes were labeled with radioactive isotopes, but modern applications use probes labeled with nonradioactive compounds that undergo chemical or color reactions to indicate the location of a specific clone in a library.

Although less central to research today, libraries still have their place in certain applications in the genetics lab. However, as you will learn later in the text (see Chapter 21), the basic methods of recombinant DNA technology, including DNA libraries, were the foundation for the development of more powerful whole-genome techniques that led to the **genomics** era of modern genetics and molecular biology. Genomic techniques, in which entire genomes are being sequenced without creating libraries, have largely replaced libraries at least for cloning and isolating one or a few genes at a time.

20.3 The Polymerase Chain Reaction Is a Powerful Technique for Copying DNA

As we will discuss later in the text (see Chapter 22), recombinant DNA techniques developed in the early 1970s gave birth to the biotechnology industry because these methods enabled scientists to clone human genes (such as the insulin gene) whose protein products could be used for therapeutic purposes. However, cloning DNA using vectors and host cells is labor intensive and time consuming. In 1986, a technique, called the **polymerase chain reaction (PCR),** was developed. PCR revolutionized recombinant DNA methodology and further accelerated the pace of biological research. The significance of this method was underscored by the awarding of the 1993 Nobel Prize in Chemistry to Kary Mullis, who developed the technique.

PCR is a rapid method of DNA cloning that extends the power of recombinant DNA and in many cases eliminates the need to use host cells for cloning. PCR is a method of choice for many applications, whether in molecular biology, human genetics, evolution, development, conservation, or forensics.

By copying a specific DNA sequence through a series of *in vitro* reactions, PCR can amplify target DNA sequences that are initially present in very small quantities in a population of other DNA molecules. When performing PCR, double-stranded target DNA to be amplified is placed in a tube with DNA polymerase, Mg^{2+} (an important cofactor for DNA polymerase), and the four deoxyribonucleoside triphosphates. In addition, some information about the nucleotide sequence of the target DNA is required. This sequence information is used to synthesize two oligonucleotide **primers:** short (typically about 20 nucleotides long) single-stranded DNA sequences, one complementary to the 5′ end of one strand of target DNA to be amplified and another primer complementary to the opposing strand of target DNA at its 3′ end.

When added to a sample of double-stranded DNA that has been denatured into single strands, the primers bind to complementary nucleotides flanking the sequence to be cloned. DNA polymerase can then extend the 3′ end of each primer to synthesize second strands of the target DNA. One complete reaction process, called a **cycle**, doubles the number of DNA molecules in the reaction [Figure 20.7(a)].

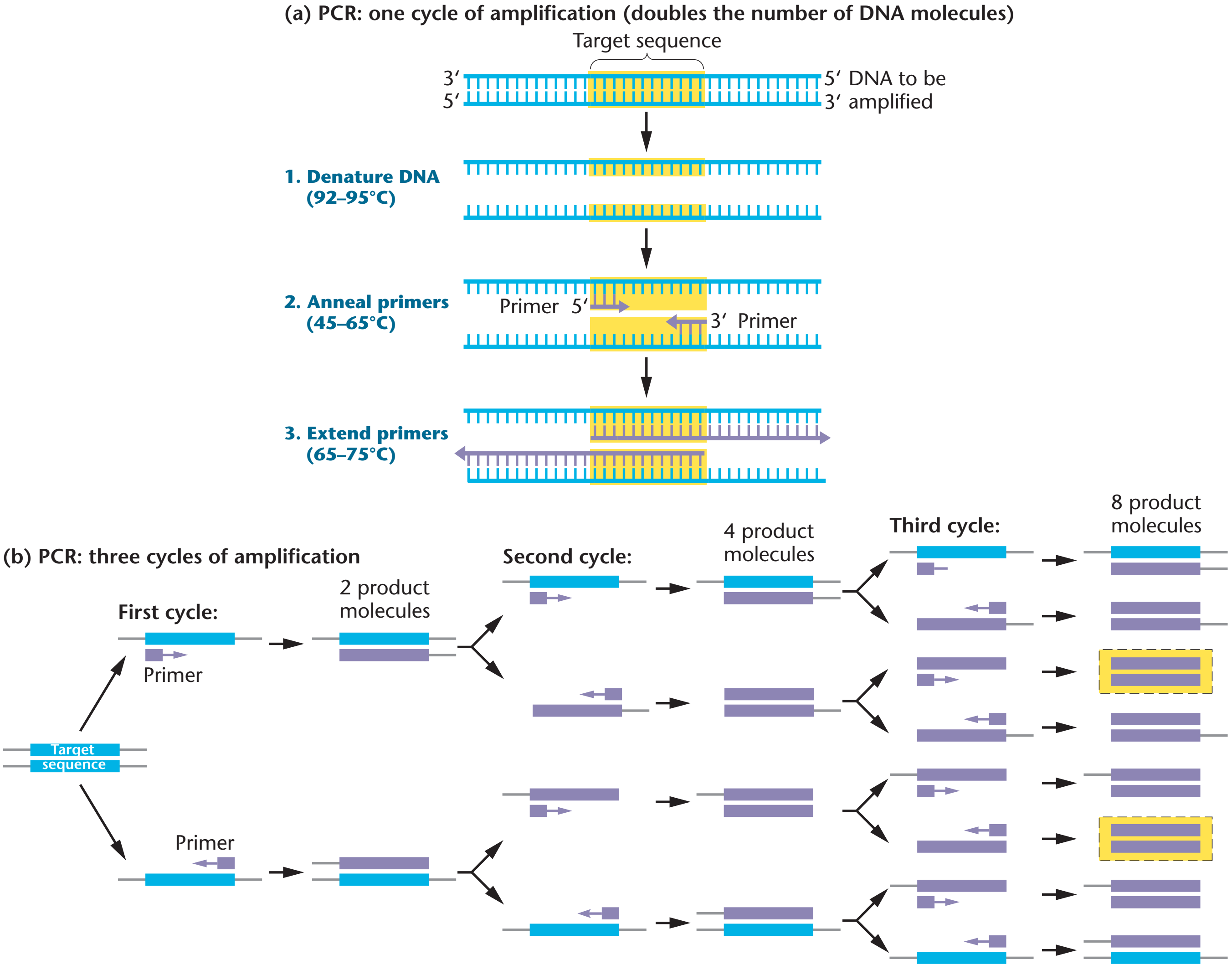

(a) PCR: one cycle of amplification (doubles the number of DNA molecules)
Target sequence
3′
5′
5′ DNA to be
3′ amplified
1. Denature DNA (92–95°C)
2. Anneal primers (45–65°C)
Primer 5′
3′ Primer
3. Extend primers (65–75°C)
(b) PCR: three cycles of amplification
First cycle:
Primer
Target sequence
Primer
2 product molecules
Second cycle:
4 product molecules
Third cycle:
8 product molecules

FIGURE 20.7 Steps in the polymerase chain reaction (PCR). (a) In this schematic representation, a relatively short sequence of DNA is shown being amplified. Typically, the segments of DNA used for PCR are several thousand nucleotides in length, and the primers bind somewhere within the actual DNA molecule and not so close to the ends as in the schematic. Notice that the first cycle produces amplified molecules with a strand that extends beyond the target sequence. (b) Repeated cycles of PCR can quickly amplify the target DNA sequence more than a millionfold. Products in part (b) that consist of only the target sequence are outlined and highlighted.

Repetition of the process produces large numbers of copied target DNA very quickly [Figure 20.7(b)]. If desired, the PCR products can be cloned into plasmid vectors for further use.

The amount of amplified DNA produced is theoretically limited only by the number of times these cycles are repeated, although several factors prevent PCR reactions from amplifying very long stretches of DNA. Most routine PCR applications involve a series of three reaction steps in a cycle. These three steps are as follows:

1. ***Denaturation:*** The double-stranded DNA to be cloned is *denatured* into single strands by heating to 92–95°C for about 1 minute. The DNA can come from many sources, including genomic DNA, mummified remains, fossils, or forensic samples such as dried blood, or semen, or hair.
2. ***Hybridization/Annealing:*** The temperature of the reaction is lowered to between 45°C and 65°C, which causes primer binding, also called hybridization or annealing, to the denatured, single-stranded DNA. The primers serve as starting points for DNA polymerase to synthesize new DNA strands complementary to the target DNA. Factors such as primer length, base composition of primers (GC-rich primers are more thermally stable than AT-rich primers), and whether or not all bases in a primer are complementary to bases in the target sequence are among primary considerations when selecting a hybridization temperature for an experiment.
3. ***Extension:*** The reaction temperature is adjusted to between 65°C and 75°C, and DNA polymerase uses the primers as a starting point to synthesize new DNA strands by adding nucleotides to the ends of the primers in a 5′ to 3′ direction.

PCR is a “chain reaction” because it involves a chain of reactions in series where each cycle is round of amplification. As a consequence, the number of new DNA strands is doubled in each cycle, and the new strands, along with the old strands, serve as templates in the next cycle. Each cycle takes 2 to 5 minutes and can be repeated immediately, so that in less than 3 hours, 25 to 30 cycles result in over a million-fold increase in the amount of DNA. The process is automated by instruments called *thermocyclers,* or simply PCR machines, that can be programmed to carry out a predetermined number of cycles. The large amounts of a specific DNA sequence produced can be used for many purposes, including cloning into plasmid vectors, DNA sequencing, clinical diagnosis, and genetic screening.

PCR requires a special type of DNA polymerase. Multiple PCR cycles involve repetitive heating and cooling of samples, which eventually lead to heat denaturation and loss of activity of most proteins. PCR reactions rely on thermostable forms of DNA polymerase capable of withstanding multiple heating and cooling cycles without significant loss of activity. PCR became a major tool when DNA polymerase was isolated from *Thermus aquaticus,* a bacterium living in habitats like the hot springs of Yellowstone National Park, where it was first discovered. Called *Taq* polymerase, this enzyme is capable of tolerating extreme temperature changes and was the first thermostable polymerase used for PCR.

PCR-based DNA cloning has several advantages over library cloning approaches. PCR is rapid and can be carried out in a few hours, rather than the days required for making and screening DNA libraries. PCR is also very sensitive and amplifies specific DNA sequences from vanishingly small DNA samples, including the DNA in a single cell. This feature of PCR is invaluable in several kinds of applications, including genetic testing, forensics, and molecular paleontology. With carefully designed primers, DNA samples that have been partially degraded, contaminated with other materials, or embedded in a matrix (such as the fossilized tree resin known as amber) can be recovered and amplified. Frequently scientists use PCR to replicate small pieces of a DNA sequence of interest to make probes—a process often referred to as *subcloning.* A wide variety of PCR-based techniques, including techniques that are essential for studying whole genomes, involve different variations of the basic technique just described. Several commonly used variations will be discussed shortly.

PCR Limitations

Although PCR is a valuable technique, it does have limitations: one being that some information about the nucleotide sequence of the target DNA must be known in order to synthesize primers. In addition, even minor contamination of the sample with DNA from other sources can cause problems. For example, cells shed from a researcher’s skin can contaminate samples gathered from a crime scene or taken from fossils, making it difficult to obtain accurate results. PCR reactions must always be performed in parallel with carefully designed and appropriate controls. Also, PCR typically cannot amplify particularly long segments of DNA. Normally, DNA polymerase in a PCR reaction extends primers only for relatively short distances; it does not continue processively until it reaches the other end of long template strands of DNA. Because of this characteristic, scientists tend to amplify pieces of DNA that are only several thousand nucleotides in length, which is fine for most routine PCR applications.

PCR Applications

In the decades since its invention, PCR has become one of the most versatile and widely used techniques in modern genetics and molecular biology. PCR and its variations now have many other applications beyond amplification for DNA cloning, including roles in DNA sequencing and a variety of techniques for genomic analysis.

As you will learn later in the text (see Chapter 22), gene-specific primers provide a way of using PCR for screening mutations involved in genetic disorders, thus making PCR an important technique for diagnosing those disorders. PCR is also a key diagnostic methodology for detecting bacteria and viruses (such as hepatitis or HIV) in humans, and pathogenic bacteria such as *E. coli* and *Staphylococcus aureus* in contaminated food.

PCR techniques are particularly advantageous when studying samples from single cells, fossils, or a crime scene, where a single hair or even a saliva-moistened postage stamp can serve as the source of the DNA. Later in the text (see Special Topic Chapter 2—DNA Forensics), we will discuss how PCR is used in human identification, including remains identification, and in forensic applications.

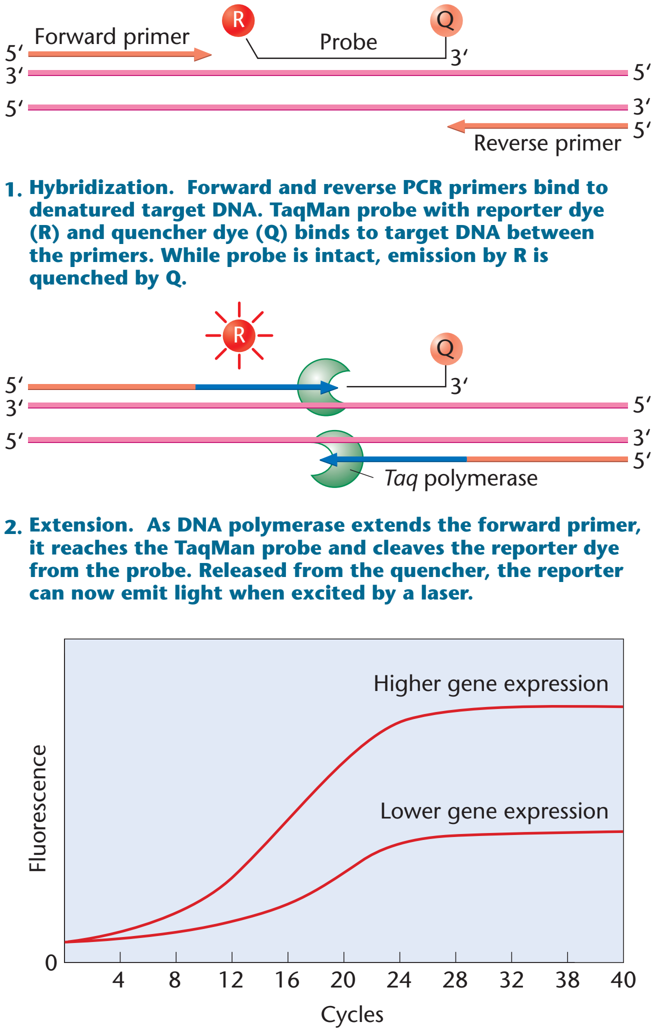
Reverse transcription PCR (RT-PCR) is a powerful methodology for studying gene expression—that is, mRNA production by cells or tissues. In RT-PCR, RNA is isolated from cells or tissues to be studied, and reverse transcriptase is used to generate double-stranded cDNA molecules, as described earlier when we discussed preparation of cDNA libraries. PCR is then used to amplify cDNA with a set of primers specific for the gene of interest. Amplified cDNA fragments are separated and visualized on an agarose gel.

Because the amount of amplified cDNA in RT-PCR is based on the relative number of mRNA molecules in the starting reaction, RT-PCR can be used to evaluate relative levels of gene expression in different samples. The amplified cDNA can be inserted into plasmid vectors, which are replicated to produce a cDNA library. In turn this library can be used to sequence the cDNA samples if desired, to make probes, or to select specific cDNAs for further analysis, among other applications. RT-PCR is more sensitive than traditional cDNA library preparation and is a powerful tool for identifying mRNAs that are not highly expressed in a cell.

One of the most valuable modern PCR techniques involves a method called **quantitative real-time PCR (qPCR)** or simply real-time PCR. This approach makes it possible to determine the amount of PCR product made during an experiment, which enables researchers to quantify amplification reactions as they occur in "real time." qPCR is a powerful and rapid technique for measuring and quantitating changes in gene expression, particularly when multiple samples and different genes are being analyzed.

There are several ways to run qPCR experiments, but the basic procedure involves the use of specialized thermal cyclers equipped with a laser to scan a beam of light through each PCR tube. Each reaction tube contains either a dye-containing probe or a DNA-binding dye, both of which emit fluorescent light when illuminated by the laser. The light emitted by these dyes correlates to the amount of amplified PCR product. Light from each tube is captured by a detector that relays information to a computer to provide a readout on the amount of fluorescence produced after each cycle, and thus the precise number of molecules in the original sample.

Two commonly used reagents for qPCR are a dye called SYBR® Green and TaqMan® probes. SYBR Green is a dye that binds double-stranded DNA. As more DNA is copied with each round of real-time PCR, there are more double-stranded DNA molecules to bind SYBR Green, which increases the amount of fluorescent light emitted. TaqMan probes are complementary to specific regions of the target DNA between where the forward and reverse primers for PCR bind. As with SYBR Green, increases in DNA copy number are reflected in increases in fluorescent light from TaqMan probes. The TaqMan approach is described in **Figure 20.8** and the accompanying figure legend.



1. Hybridization. Forward and reverse PCR primers bind to denatured target DNA. TaqMan probe with reporter dye (R) and quencher dye (Q) binds to target DNA between the primers. While probe is intact, emission by R is quenched by Q.

2. Extension. As DNA polymerase extends the forward primer, it reaches the TaqMan probe and cleaves the reporter dye from the probe. Released from the quencher, the reporter can now emit light when excited by a laser.

3. Detection. Emitted light from the reporter is detected and interpreted to produce a plot that quantitates the amount of PCR product produced with each cycle.

FIGURE 20.8 The TaqMan approach to quantitative real-time PCR (qPCR) involves a pair of PCR primers along with a probe sequence complementary to the target gene. The probe contains a reporter dye (R) at one end and a quencher dye (Q) at the other end. While the quencher dye is close to the reporter dye, it interferes with fluorescence released by the reporter dye. When *Taq* DNA polymerase extends a primer to synthesize a strand of DNA, it cleaves the reporter dye off of the probe allowing the reporter to give off energy. Each subsequent PCR cycle releases more reporter dyes. A readout of the resulting increase in fluorescence is produced by a computer.

NOW SOLVE THIS

20.2 You have just created the world's first genomic library from the African okapi, a relative of the giraffe. No genes from this genome have been previously isolated or described. You wish to isolate the gene encoding the oxygen-transporting protein β-globin from the okapi library. This gene has been isolated from humans, and its nucleotide sequence and amino acid sequence are available in databases. Using the information available about the human β-globin gene, what strategy can you use to isolate this gene from the okapi library?

- **HINT:** *This problem asks you to design PCR primers to amplify the β-globin gene from a species whose genome you just sequenced. The key to its solution is to remember that you have at your disposal sequence data for the human β-globin gene and to consider that PCR experiments require the use of primers that bind to complementary bases in the DNA to be amplified.*

For more practice, see Problems 14 and 16.

20.4 Molecular Techniques for Analyzing DNA and RNA

A wide range of molecular techniques is available to almost anyone who does research involving DNA and RNA, particularly those who study the structure, expression, and regulation of genes. There are far too many techniques available to modern geneticists than we can address in this chapter. In the following sections, we consider some of the techniques that are most routinely used to analyze DNA and RNA. Throughout later sections of the text you will see these and other techniques discussed in the context of certain applications in modern genetics.

Restriction Mapping

Historically, one of the first steps in characterizing a DNA clone was the construction of a **restriction map.** A restriction map reports the number of, order of, and distances between restriction-enzyme cleavage sites along a cloned segment of DNA, thus providing information about the length of the cloned insert and the location of restriction-enzyme cleavage sites within it. This information could be used to subclone fragments of a gene or compare its organization with that of other cloned sequences. In the Human Genome Project, restriction maps of the human genome were key to digesting the genome into pieces that could be sequenced.

Before DNA sequencing and bioinformatics became popular, restriction maps were created by cutting DNA with different restriction enzymes and separating DNA fragments by gel electrophoresis, a method that separates fragments by size, with the smallest pieces moving farthest through the gel (see Chapter 10; refer to Figure 10.18). The fragments form a series of bands that can be visualized by staining with DNA-binding stains such as *ethidium bromide* and illuminating it with ultraviolet light (**Figure 20.9**). The digestion pattern of fragments generated could then be interpreted to determine the location of restriction sites for different enzymes.

While agarose gel electrophoresis remains a routine and essential technique in the laboratory, because of advances in DNA sequencing and the use of bioinformatics, restriction maps are now almost always created by simply using software to identify restriction-enzyme cutting sites in sequenced DNA. The Exploring Genomics exercise in this chapter involves a Web site, Webcutter, which is commonly used for generating restriction maps.

Nucleic Acid Blotting

Several of the techniques described in this chapter and elsewhere in the book rely on hybridization between complementary nucleic acid (DNA or RNA) molecules. One of the most widely used methods for detecting such hybrids is called **Southern blot analysis** or simply **Southern blotting** (after Edwin Southern, who devised it). Southern blotting is another pioneering method that served essential roles in the early decades of DNA cloning such as identifying which clones in a library contained a given DNA sequence, identifying specific genes in genomic DNA digested with a restriction enzyme, and identifying the number of copies of a particular sequence or gene that are present in a genome. Most all of these application examples have now been replaced by modern DNA sequencing approaches.

Southern blotting has two components: separation of DNA fragments by gel electrophoresis, and hybridization of the fragments using labeled probes. Gel electrophoresis can be used to characterize the number of fragments produced by restriction

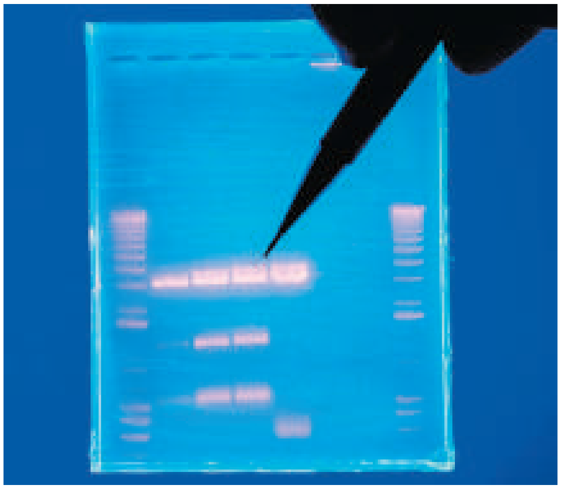


FIGURE 20.9 An agarose gel containing separated DNA fragments stained with a DNA-binding dye (ethidium bromide) and visualized under ultraviolet light. Smaller fragments migrate faster and farther than do larger fragments, resulting in the distribution shown. Molecular techniques involving agarose gel electrophoresis are routinely used in a wide range of applications.

digestion and to estimate their molecular weights, when the number of fragments generated is relatively low. However, restriction enzyme digestion of large genomes—such as the human genome, with more than 3 billion nucleotides—would produce so many different fragments that they would run together on a gel to produce a continuous smear. The identification of specific fragments in these cases is accomplished in the next step: hybridization characterizes the DNA sequences present in the fragments. The steps involved in the complete Southern blotting technique are shown in **Figure 20.10**.

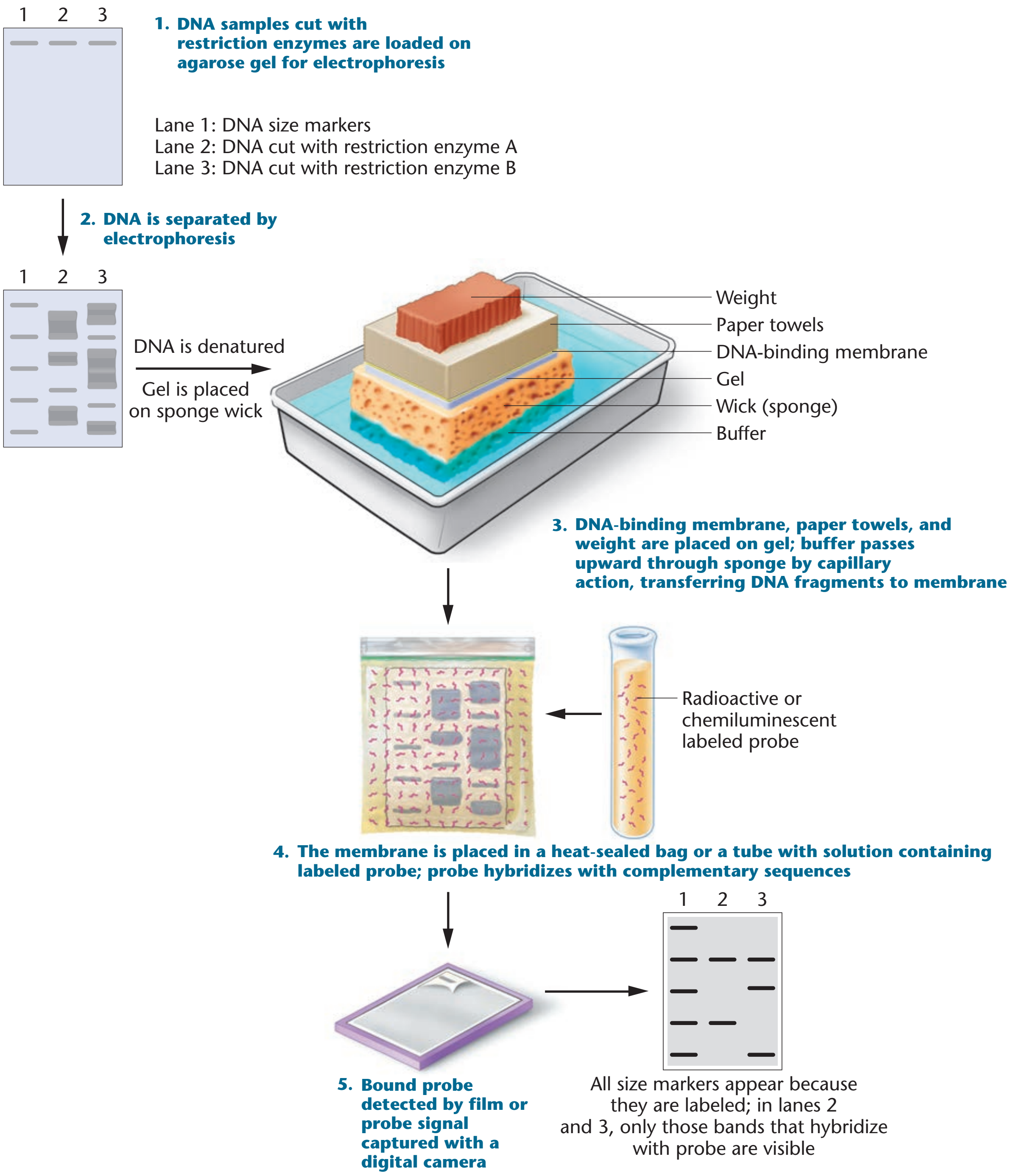

1 2 3
1. DNA samples cut with restriction enzymes are loaded on agarose gel for electrophoresis
Lane 1: DNA size markers
Lane 2: DNA cut with restriction enzyme A
Lane 3: DNA cut with restriction enzyme B
2. DNA is separated by electrophoresis
1 2 3
DNA is denatured
Gel is placed on sponge wick
Weight
Paper towels
DNA-binding membrane
Gel
Wick (sponge)
Buffer
3. DNA-binding membrane, paper towels, and weight are placed on gel; buffer passes upward through sponge by capillary action, transferring DNA fragments to membrane
Radioactive or chemiluminescent labeled probe
4. The membrane is placed in a heat-sealed bag or a tube with solution containing labeled probe; probe hybridizes with complementary sequences
1 2 3
5. Bound probe detected by film or probe signal captured with a digital camera
All size markers appear because they are labeled; in lanes 2 and 3, only those bands that hybridize with probe are visible

FIGURE 20.10 In the Southern blotting technique, samples of the DNA to be analyzed are digested with restriction enzymes and the fragments are separated by gel electrophoresis. The pattern of fragments in the gel is visualized and photographed under ultraviolet illumination. The DNA fragments in the gel are denatured by soaking the gel in an alkaline solution. Then the gel is placed on a sponge wick that is in contact with a buffer solution and covered with a DNA-binding membrane. Layers of paper towels or blotting paper are placed on top of the membrane and held in place with a weight. Capillary action draws the buffer through the gel, transferring the pattern of DNA fragments from the gel to the membrane. DNA fragments are fixed onto the membrane and then hybridized to a labeled DNA probe. The membrane is washed to remove excess probe and overlaid with a piece of X-ray film for autoradiography or detected with a digital camera with chemiluminescence probes. Only fragments hybridizing to the probe are visualized.

To produce Figure 20.11, researchers cut samples of genomic DNA with several restriction enzymes. The pattern of fragments obtained for each restriction enzyme is shown in Figure 20.11(a). A Southern blot of this gel is illustrated in Figure 20.11(b). The probe hybridized to complementary sequences, identifying fragments of interest.

Southern blotting led to the subsequent development of other widely used blotting approaches. RNA blotting was called **Northern blot analysis** or simply **Northern blotting,** and, following a naming scheme that correlates with the directionality of a compass, a related blotting technique involving proteins is known as **Western blotting.** Western blotting is a routine technique for analyzing proteins. Thus part of the historical significance of Southern blotting is that it led to the development of other blotting methods that are key tools for studying nucleic acids and proteins.

Prior to the development of RT-PCR and real-time PCR, Northern blotting was commonly used to study gene expression. To determine whether a gene is actively being expressed in a given cell or tissue type, Northern blotting probes for the presence of mRNA complementary to a cloned gene (Figure 20.12). To do this, mRNA is extracted from a specific cell or tissue type and separated by gel electrophoresis. The resulting pattern of RNA bands is transferred to a membrane, as in Southern blotting. The membrane is then exposed to a labeled single-stranded DNA or RNA probe derived from a cloned copy of the gene. If mRNA complementary to the DNA probe is present, the complementary sequences will hybridize and be detected. Northern blots provide information about the expression of specific genes and are used to study patterns of gene expression in embryonic tissues, cancer, and genetic disorders.

As Figure 20.12 shows, Northern blotting can be used to derive the size of a gene's mRNA transcripts and to flag the possibility of alternative splicing. Measuring band density gives an estimate of the relative transcriptional activity of the gene. Thus, Northern blots both characterize and quantify the transcriptional activity of genes. Northern blots are occasionally still used to study RNA expression, but because PCR-based techniques are faster and more sensitive than blotting methods, techniques such as RT-PCR are often the preferred approach, particularly for measuring changes in gene expression.

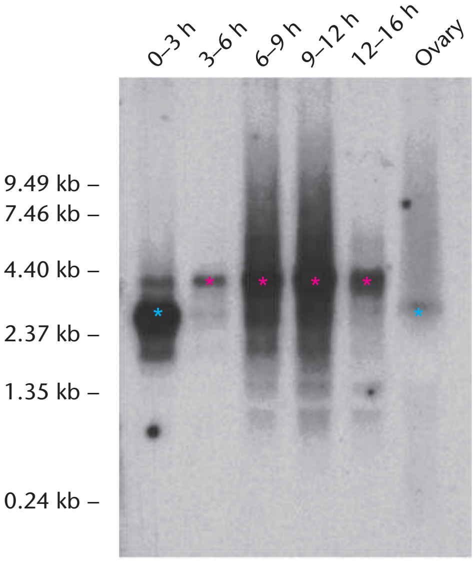


FIGURE 20.12 Northern blot analysis of *dFmr1* gene expression in *Drosophila* during embryogenesis. A prominent *dfmr1* RNA transcript of approximately 2.8 kb is present in ovaries and 0- to 3-hour-old embryos (blue asterisks). At 3 to 6 hours and beyond, the major transcript detected is about 4.0 kb and peaks in abundance between 9 and 12 hours of embryonic development (red asterisks). Note, transcripts of different lengths (4.0, 2.8 kb) are produced from the same gene, and that the levels of expression for each transcript vary during different stages of development. These data suggest that *dFmr1* gene expression may be regulated at the levels of transcription or transcript processing during embryogenesis. The *dFmr1* gene is a homolog of the human *FMR1* gene. Loss-of-function mutations in *FMR1* result in human Fragile-X mental retardation.

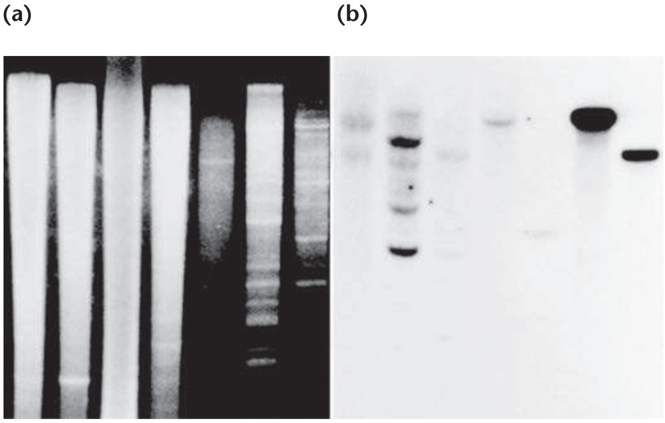


FIGURE 20.11 (a) Agarose gel stained with ethidium bromide to show DNA fragments. (b) Chemiluminescent image of a Southern blot prepared from the gel in part (a). Only those bands containing DNA sequences complementary to the probe show hybridization.

In Situ Hybridization

Finally, as noted earlier in the text (see Chapter 10), **fluorescence *in situ* hybridization (FISH)** is a powerful tool that involves hybridizing a probe directly to a chromosome or RNA without blotting (see Figure 10.17, Chapter 17 opening photograph, and Figures 23.9 and 23.10). FISH can be carried out with isolated chromosomes on a slide or directly *in situ* in tissue sections or entire organisms, particularly when embryos are used for various studies in developmental genetics. For example, in developmental studies one can identify which cell types in an embryo

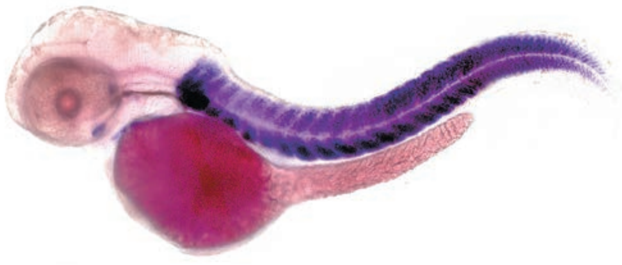


FIGURE 20.13 *In situ* hybridization of a zebrafish embryo 48 hours after fertilization showing expression of *atp2a1* mRNA, which encodes a muscle-specific calcium pump. The probe revealing *atp2a1* expression produces dark blue staining. Notice that this staining is restricted to muscle cells surrounding the developing spinal cord of the embryo.

express different genes during specific stages of development (Figure 20.13).

Variations of the FISH technique are also being used to produce **spectral karyotypes** in which individual chromosomes can be detected using probes labeled with dyes that will fluoresce at different wavelengths (see Chapter 8 opening photograph and Figure 24.1). Spectral karyotyping has proven to be extremely valuable for detecting deletions, translocations, duplications, and other anomalies in chromosome structure, such as chromosomal rearrangements (discussed in Chapter 8), and for detecting chromosomal abnormalities in cancer cells (discussed in Chapter 24).

20.5 DNA Sequencing Is the Ultimate Way to Characterize DNA at the Molecular Level

In a sense, cloned DNA, from a single gene to an entire genome, is completely characterized at the molecular level only when its nucleotide sequence is known. The ability to sequence DNA has greatly enhanced our understanding of genome organization and increased our knowledge of gene structure, function, and mechanisms of regulation.

Historically, the most commonly used method of DNA sequencing was developed by Fred Sanger and colleagues during the 1970s and is known as **dideoxy chain-termination sequencing** or simply **Sanger sequencing.** Because Sanger sequencing was an important foundational method for newer, more modern approaches to DNA sequencing, we will briefly discuss this approach here. In this technique, a double-stranded DNA molecule whose sequence is to be determined is converted to single strands that are used as a template for synthesizing a series of complementary strands. The DNA to be sequenced is mixed with a primer that is complementary to the target DNA or vector, along with DNA polymerase, and the four deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP) are added to each tube.

The key to the Sanger technique is the addition of a small amount of modified deoxyribonucleotides, called **dideoxynucleotides** (abbreviated ddNTPs) (**Figure 20.14,** inset box). Notice that a dideoxynucleotide has a 3′ hydrogen instead of a 3′ hydroxyl group. Dideoxynucleotides are called chain-termination nucleotides because they lack the 3′ oxygen required to form a phosphodiester bond with another nucleotide. If ddNTPs are included in a DNA synthesis reaction, the polymerase will occasionally (randomly) insert a ddNTP instead of a dNTP into a growing DNA strand. Once this occurs, synthesis terminates because DNA polymerase cannot add new nucleotides to a ddNTP due to its lack of a 3′ oxygen. The Sanger reaction takes advantage of this key modification.

For example, in Step 2 of Figure 20.14, notice that the shortest fragment generated is a sequence that has added ddCTP to the 3′ end of the primer and the chain has terminated. Over time as the reaction proceeds, eventually a ddNTP will be inserted at every location in the sequence of newly synthesized DNA molecules so that each strand synthesized differs in length by one nucleotide and is terminated by a ddNTP. This allows for separation of reaction products by gel electrophoresis, which can then be used to determine the sequence.

When the Sanger technique was first developed, four separate reaction tubes, each with a different single ddNTP (e.g., ddATP, ddCTP, ddGTP, and ddTTP), were used. These reactions typically used either a radioactively labeled primer or a radioactively labeled ddNTP to permit analysis of the sequence following polyacrylamide gel electrophoresis and autoradiography. Historically, this approach involved large polyacrylamide gels in which each reaction was loaded on a separate lane of the gel and ladder-like banding patterns revealed by autoradiography were read to determine the sequence. This original approach could typically read about 800 bases of 100 DNA molecules simultaneously. *Read length*—that is, the amount of sequence that can be generated in a single individual reaction—and the total amount of DNA sequence generated in a sequence *run* (which is effectively read length times the number of reactions an instrument can run during a given period of time) together have become a hot area for innovation in sequencing technology.

Modifications of the Sanger technique in the mid-1980s led to technologies that allowed sequencing reactions to occur in a single tube. As shown in Figure 20.14, each of the four ddNTPs were labeled with a different-colored fluorescent dye. These reactions were carried out in PCR-like fashion using cycling reactions that permit

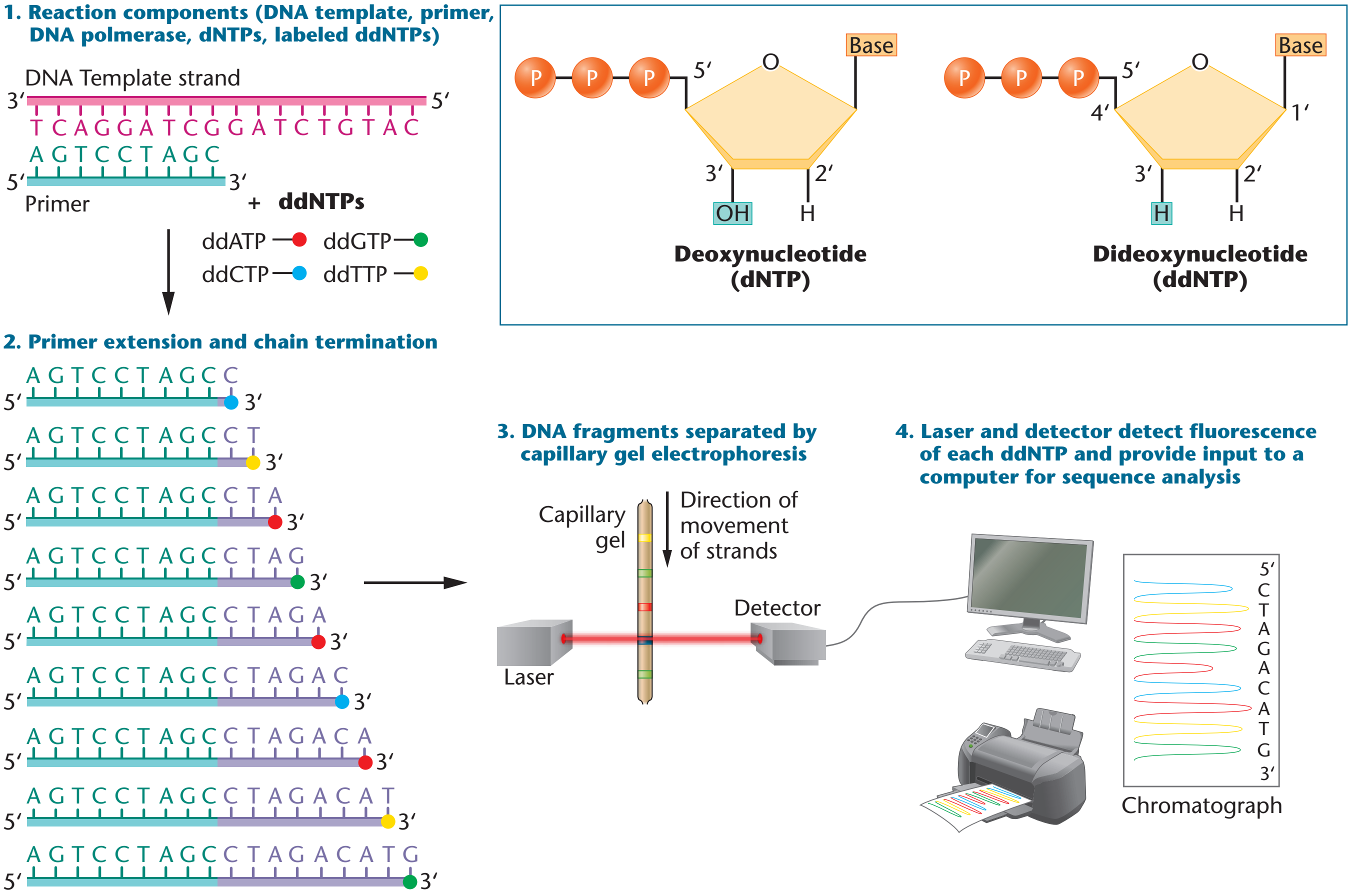

1. Reaction components (DNA template, primer, DNA polmerase, dNTPs, labeled ddNTPs)
DNA Template strand
3′ T C A G G A T C G G A T C T G T A C 5′
A G T C C T A G C
5′ 3′
Primer
+ ddNTPs
ddATP
ddGTP
ddCTP
ddTTP
P P P 5′ O Base
3′ 2′
OH H
Deoxynucleotide (dNTP)
P P P 5′ O Base
4′ 1′
3′ 2′
H H
Dideoxynucleotide (ddNTP)
2. Primer extension and chain termination
A G T C C T A G C C
A G T C C T A G C C T
A G T C C T A G C C T A
A G T C C T A G C C T A G
A G T C C T A G C C T A G A
A G T C C T A G C C T A G A C
A G T C C T A G C C T A G A C A
A G T C C T A G C C T A G A C A T
A G T C C T A G C C T A G A C A T G
5′
3′
3. DNA fragments separated by capillary gel electrophoresis
Capillary gel
Direction of movement of strands
Laser
Detector
4. Laser and detector detect fluorescence of each ddNTP and provide input to a computer for sequence analysis
5′ C T A G A C A T G 3′
Chromatograph

FIGURE 20.14 Computer-automated DNA sequencing using the chain-termination (modified Sanger) method. The inset box at upper right illustrates dideoxynucleotide (ddNTP) structure. (1) A primer is annealed to a sequence adjacent to the DNA being sequenced (usually near the multiple cloning site of a cloning vector). A reaction mixture is added to the primer–template combination. This includes DNA polymerase, the four dNTPs, and small molar amounts of ddNTPs labeled with fluorescent dyes. (2) All four ddNTPs are added to the same reaction tube. During primer extension, the polymerase occasionally (randomly) inserts a ddNTP instead of a dNTP, terminating the synthesis of the chain because the ddNTP does not have the OH group needed to attach the next nucleotide. Over the course of the reaction, all possible termination sites will have a ddNTP inserted, thus all possible lengths of chains are produced. The products of the reaction are added to a single lane on a capillary gel (3), and the bands are read by a detector and imaging system (4). from the newly synthesized strand. In this case, the sequence obtained begins with 5′-CTAGACATG-3′,as seen in the chromatograph in step 4.

greater read and run capabilities. The reaction products were separated through a single, ultrathin-diameter polyacrylamide tube gel called a capillary gel (*capillary gel electrophoresis*). As DNA fragments move through the gel, they are scanned with a laser. The laser stimulates fluorescent dyes on each DNA fragment, which then emit different wavelengths of light for each ddNTP. Emitted light is captured by a detector that amplifies and feeds this information into a computer to convert the light patterns into a DNA sequence that is technically called an electropherogram or chromatograph. The data are represented as a series of colored peaks, each corresponding to one nucleotide in the sequence.

For about two decades following the early 1990s, DNA sequencing was largely performed through computer-automated Sanger-reaction-based technology (shown in Figure 20.14) and referred to as **computer-automated high-throughput DNA sequencing.** These systems were a big improvement over manual Sanger systems because they could generate relatively large amounts of DNA sequence in relatively short periods of time. Computer-automated sequences could achieve read lengths of approximately 1000 bp with about 99.999 percent accuracy for about $0.50 per kb. Automated DNA sequencers of this time period often contained multiple capillary gels (as many as 96) that were several feet long and could process several thousand bases of sequences, so many of these instruments made it possible to generate over 2 million bp of sequences in a day! Such systems became essential for the rapidly accelerating progress of the Human Genome Project. But by around 2005 and in the time since, sequencing technologies were to improve dramatically.

Sequencing Technologies Have Progressed Rapidly

Fred Sanger was awarded part of the 1980 Nobel Prize in Chemistry (which he shared with Walter Gilbert and Paul Berg) for contributions to sequencing technology. In the nearly four decades since then, DNA sequencing technologies have undergone an incredible evolution leading to dramatic improvements in sequencing capabilities. New innovations in sequencing technology have been developing especially quickly in the past decade.

Sanger sequencing approaches, including those involving computer-automated instruments, are rarely used today except for occasional sequencing of a relatively short piece of DNA or in labs that cannot afford more expensive sequencing instruments. When it comes to sequencing entire genomes, Sanger sequencing technologies and early-generation computer-automated approaches are outdated. Compared to newer technologies, the costs of those approaches were relatively high, and sequencing output, even with computer automation, was simply not high enough to support the growing demand for genomic data. This demand is being driven in large part by *personalized genomics* (see Chapter 21) and the desire to reveal the genetic basis of human diseases, which involves routine sequencing of complete individual human genomes.

Next-Generation Sequencing Technology

The development of genomics has spurred a demand for sequencers that are capable of generating millions of bases of DNA sequences in a relatively short time. As we will discuss later in the text (see Chapter 21), in the mid-2000s a race was on to develop sequencing technologies that would allow the complete sequencing of an individual human genome for $1000. **Next-generation sequencing (NGS) technologies** (the second generation after Sanger methods) were the next big advance in DNA sequencing and a platform many thought would achieve the $1000 genome goal.

NGS technologies dispensed with *first-generation* methods (the Sanger technique and capillary electrophoresis) in favor of sophisticated, parallel formats (simultaneous reaction formats) that synthesized DNA from tens of thousands of identical strands simultaneously and then used state-of-the-art fluorescence imaging techniques to detect new synthesized strands and average sequence data across many molecules being sequenced. NGS technologies provided an unprecedented capacity for generating massive amounts of DNA sequence data rapidly and at dramatically reduced costs per base. NGS has become so routine that now when scientists talk about "sequencing" we are referring to NGS.

The desire for next-generation sequencing within the research community and challenges such as the $1000 genome led to an intense race among companies eager to produce NGS methods. As a result, by 2005, several NGS technologies had emerged. Some of the first instruments were capable of producing as much data as 50 capillary electrophoresis systems and were up to 200 times faster and cheaper than conventional Sanger approaches. NGS instruments generally produce short read lengths of ~50–400 bp, and then these snippets of sequence are stitched together using software to produce a coherent, complete genome.

In 2005, 454 Life Sciences was the first company to commercialize NGS technology. This approach uses a solid-phase method in which beads are attached to fragmented genomic DNA, which is then PCR amplified in separate water droplets in oil for each bead and loaded into multiwell plates and mixed with DNA polymerase. Multiwell plates often contain more than one million wells with one bead per well—each serving as a reaction tube for sequencing. Next, a sequencing technology called **pyrosequencing** is used to sequence DNA on the beads in each well. In pyrosequencing, a single, labeled nucleotide (e.g., dATP) is flowed over the wells. Each well contains a single bead along with primers annealed to the DNA on the bead. When a complementary nucleotide crosses a template strand adjacent to the primer, it is added to the 3′ end of the primer by DNA polymerase.

Incorporation of a nucleotide results in the release of pyrophosphate, which initiates a series of chemiluminescent (light-releasing) reactions that ultimately produce light using the firefly enzyme luciferase. Emitted light from the reaction is captured and recorded to determine when a single nucleotide has been incorporated into a strand. By rapidly repeating the nucleotide flow step with each of the four nucleotides to determine which base is next in the sequence, this approach can generate read lengths of about 400 bases and on the order of 500 million bases (Mbp) of data per 10-hour run.

Ultimately, several companies emerged as winners in the race to commercialize NGS technology. The Illumina® HiSeq, and variations of this instrument, is one of the most common NGS platforms currently used in cutting-edge sequencing laboratories today. This system uses a **sequencing-by-synthesis (SBS)** approach, which involves using DNA fragments as templates for synthesizing new strands, incorporating nucleotides labeled with different dyes, washing away unincorporated nucleotides, imaging incorporated nucleotides, and repeating this cycle. In the method developed by Illumina, the DNA fragments are attached to a solid support and then, using reactions similar to the Sanger method, the fluorescently tagged terminator nucleotides are added and detected. The fluorescent tags and terminator portions of the nucleotides are removed to allow another cycle of extension.

SBS methods can now generate about 600 Gb of data in 10 days—enough to sequence four complete human genomes, with each base sequenced an average of 30 times for accuracy! The instrumentation needed to run these platforms is expensive. For example, the Illumina HiSeq instrument costs about $650,000. But given the massive amounts of sequence data NGS methods can generate, the average cost per base is much lower than Sanger sequencing. Incidentally, NGS sequencing technologies also created major data management challenges for saving and storing large data files.

Third-Generation Sequencing Technology

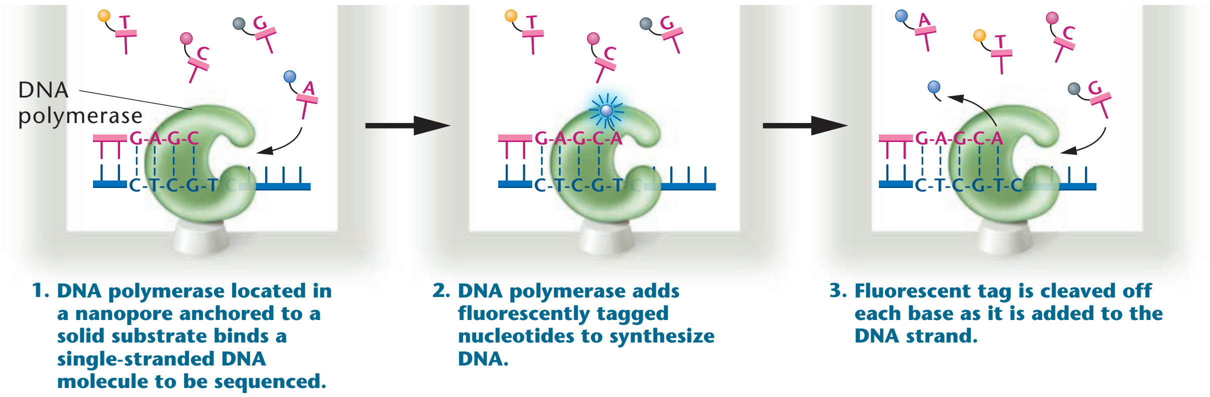
Shortly after NGS methods were commercialized, companies were announcing progress on **third-generation sequencing (TGS).** TGS methods are based on strategies that sequence a *single molecule* of single-stranded DNA, and at least four different approaches are being explored.

Pacific Biosciences' PacBio® technology is one of the leaders in TGS and involves an approach known as *single-molecule sequencing in real time (SMRT)*. The PacBio instrument works by attaching single-stranded molecules of the DNA to be sequenced to a single molecule of DNA polymerase anchored to a substrate and then visualizing, in real time, the polymerase as it synthesizes a strand of DNA (see **Figure 20.15**). The DNA polymerase is confined within a nanopore—a hole of about 10 nm in diameter located within a thin layer of metal on a glass substrate. This setup allows for the illumination necessary to detect the addition of individual nucleotides as they are added to the growing strand by the polymerase. The nucleotides are tagged with a fluorescent dye. Because the dye is attached to the terminal phosphate of the nucleotide, it is cleaved—and flashes—when its nucleotide is incorporated into the new DNA strand. Each base flashes with a characteristic color that is detected and recorded.

The PacBio was one of the first "long-read" instruments to reach the market. It generates read lengths of over 10,000 bp. In 2010, the PacBio made it possible to sequence the genomes of five strains of *Vibrio cholera* involved in a cholera outbreak in Haiti in less than an hour. Thanks to this genetic determination of the *Vibrio* strains, health workers were able to rapidly treat the outbreak with antibiotics to which these bacteria were not resistant. Despite such successes, most TGS technologies still have somewhat high error rates for sequencing accuracy—about 15 percent errors per sequence generated. The list price for one PacBio sequencer is about $350,000, which is making this technology more affordable at least for biotechnology companies, pharmaceutical companies, and well-funded academic research laboratories.

A few years ago, Oxford Nanopore Technologies developed a portable, single-molecule sequencer called the MinION that is the size of a USB memory stick! Although accuracy of this sequencer limit its applications, there is no reason to think that the technology for highly accurate pocket-sized sequencers will not advance in the near future.

We have briefly presented TGS to provide a context for how far sequencing technologies have developed since Sanger sequencing and even NGS. The detailed methodologies of TGS are less important than the conceptual significance of these approaches—a shift toward sequencing individual molecules faster, less expensively, and more accurately than previous methods. There is no doubt that the development of TGS technologies represents another significant milestone for genomic studies.



1. DNA polymerase located in a nanopore anchored to a solid substrate binds a single-stranded DNA molecule to be sequenced.
2. DNA polymerase adds fluorescently tagged nucleotides to synthesize DNA.
3. Fluorescent tag is cleaved off each base as it is added to the DNA strand.

FIGURE 20.15 Third-generation sequencing (TGS). A simplified version of one approach to TGS is shown here. In this example, a DNA polymerase molecule anchored within a nanopore binds to a single strand of DNA. As the polymerase incorporates fluorescently labeled nucleotides into a new DNA strand (shown in pink), each base emits a characteristic color that can be detected.

DNA Sequencing and Genomics

The genomics research community has embraced NGS and TGS technologies. Which approaches will eventually emerge as the sequencing methods of choice for the long term and what the next new technology will be is unclear. What is clear, however, is that the sequencing landscape has rapidly changed for the better and never before have scientists had the ability to generate so much sequence data so quickly.

Moore's law, stated by Intel co-founder Gordon Moore, is that computing power tends to double (with its price effectively halving) every 2 years. For about 50 years this has been generally true in the computing world. Scientists have often applied Moore's law to discussion about the costs of DNA sequencing. Through 2006, new technologies were cutting sequencing costs in half about every two years and keeping pace with Moore's law (**Figure 20.16**). But since 2007 the price of sequencing a human genome has plummeted dramatically, far outpacing Moore's law, in large part due to the impact of NGS entering the market.

In 2014, Illumina laid claim to the coveted $1000 genome with its HiSeq X Ten system. Although a single run would cost nearly $13,000 to generate 16 Gb of sequence data, Illumina's genome cost was calculated based on the costs savings of running ten of its systems and sequencing output over a decade of use. Other companies claim cheaper short-term costs for sequencing a genome. Now Illumina has its sights set on sequencing an entire genome for less than $100.

Our overview of DNA sequencing in this chapter is a great introduction to upcoming detailed discussions of genomics and many related topics (see Chapter 21). The principal techniques of recombinant DNA technology, particularly DNA cloning, were essential for making genome projects possible. But no technology has had a greater influence on our ability to study genomes than DNA sequencing.

Rapid advances in sequencing technology were driven by the demands of genome scientists (particularly those working on the Human Genome Project) to rapidly generate more sequence with greater accuracy and at lower cost. Because of these advances in sequencing technologies, innovative approaches to genome sequencing are driving a range of new research and clinical applications. Also (in Chapters 21 and 22) we will discuss how **RNA sequencing** has emerged as a new technique that makes it possible to measure gene expression on a genome-wide scale.

20.6 Creating Knockout and Transgenic Organisms for Studying Gene Function

Thus far we have focused on approaches to working with recombinant DNA *in vitro*. Recombinant DNA technology has also made it possible to directly manipulate genes *in vivo* in ways that allow scientists to learn more about gene function *in living organisms*. These approaches also enable scientists to create genetically engineered plants and animals for research and for commercial applications. We conclude this chapter by first briefly discussing gene knockout technology and creating transgenic animals as examples of **gene targeting**—a collection of methods that have revolutionized research in genetics. (Throughout the book, the Modern Approaches to Understanding Gene Function boxes have highlighted examples of specific research

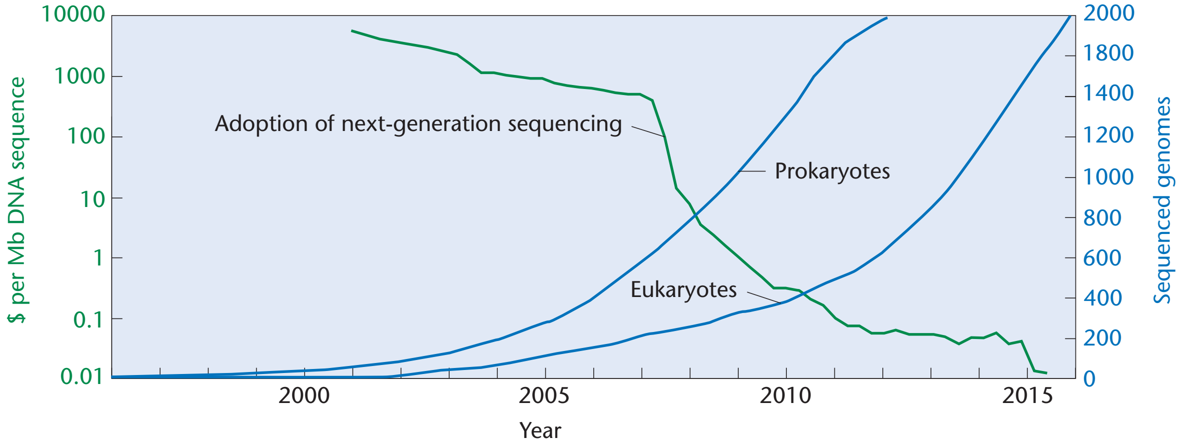


FIGURE 20.16 Rapidly decreasing costs have enabled the rapid expansion of eukaryotic and prokaryotic genome sequencing. The number of eukaryotic and prokaryotic genomes sequenced (blue lines) has increased dramatically during the past decade due to significant improvements in high-throughput sequencing technologies, particularly the adoption of next-generation sequencing, and sharply decreasing costs (green line) of sequencing large numbers of base pairs.

projects involving gene-targeting approaches.) Finally, we will look at **gene editing,** which involves the use of specific engineered enzymes to precisely modify a particular gene.

Gene Targeting and Knockout Animal Models

The concept behind gene targeting is to manipulate a specific allele, locus, or base sequence to learn about the functions of a gene of interest. Gene targeting focuses on manipulating a specific gene in the genome through approaches that involve homologous recombination. In the 1980s, scientists devised a gene-targeting technique for creating **gene knockout** (often abbreviated as KO) organisms, specifically mice. The pioneers of knockout technology, Dr. Mario Capecchi of the University of Utah and colleagues Oliver Smithies of the University of North Carolina, Chapel Hill, and Sir Martin Evans of Cardiff University, United Kingdom, received the 2007 Nobel Prize in Physiology or Medicine for developing this technique.

The fundamental purpose of creating a knockout is to *disrupt or eliminate* copies of specific gene or genes of interest and then ask, "What happens?" If physical, behavioral, and biochemical changes or other metabolic phenotypes or functions are observed in the KO animal, then one can begin to see that the gene of interest has some functional role or roles in the observed phenotypes. A knockout is an example of a **loss-of-function mutation.** Thus one of the most valuable reasons for creating a KO is to learn about gene function. The KO techniques developed in mice led to similar technologies for making KOs in zebrafish, rats, pigs, fruit flies, and many other organisms. Making KO plants is also a very active area of research.

Knockout mice have revolutionized research in genetics, molecular biology, and biomedical research in many ways. Scientists have used KO methods to create thousands of KO organisms that have advanced our understanding of gene function, animal models for many human diseases, and transgenic animals (which we will discuss on p. 477). Applications of KO technology have also provided the foundation for gene-targeting approaches in gene therapy that we discuss later in the text (see Special Topic Chapter 5—Gene Therapy).

Generally, generating a KO mouse or a transgenic mouse is a very labor intensive project that can take several years of experiments and crosses and a significant budget to complete. However, once a KO mouse is made, assuming it is fertile, a colony of mice can be maintained; often KO mice are shared around the world so that other researchers can work with them. Many companies will produce KO mice for researchers. It is also possible to make *double-knockout animals (DKOs)* and even *triple-knockout animals (TKOs).* This approach is typically used when scientists want to study the functional effects of disrupting two or three genes thought to be involved in a related mechanism or pathway.

A KO animal can be made in several ways. Here we outline a strategy for making KO mice (**Figure 20.17**), but the same basic methods apply when making most KO animals.

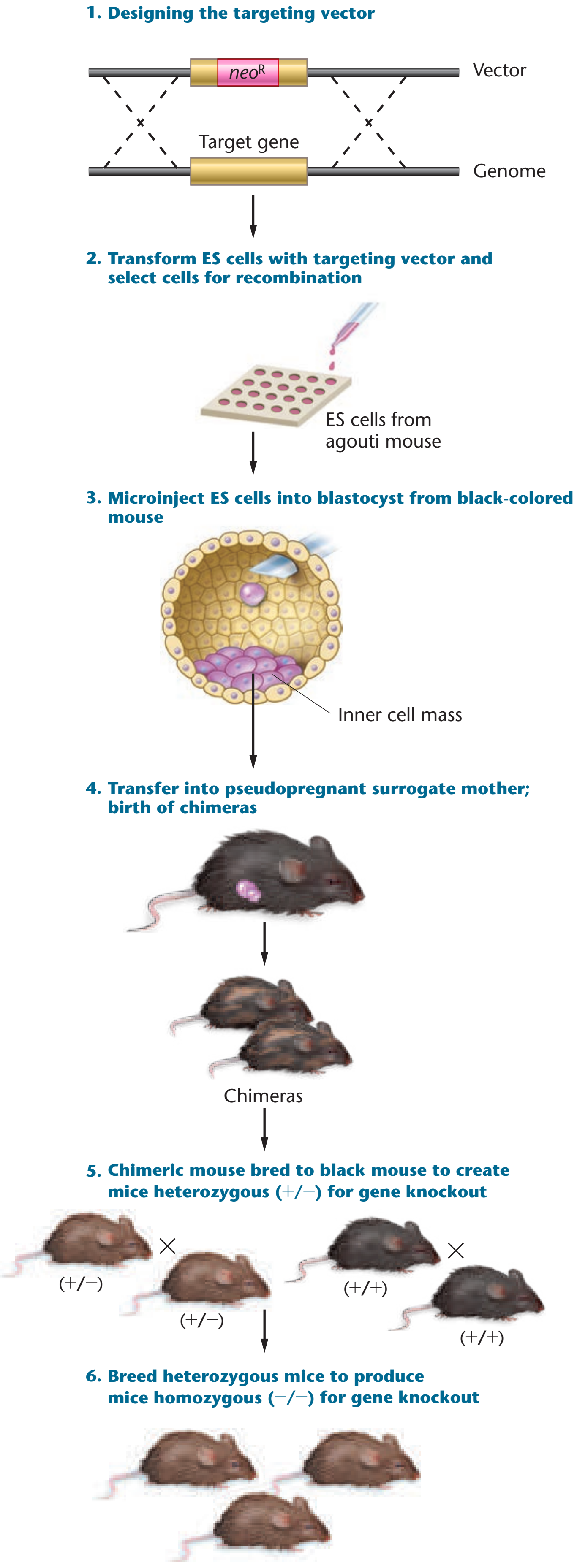

1. Designing the targeting vector
neo^R
Vector
Target gene
Genome
2. Transform ES cells with targeting vector and select cells for recombination
ES cells from agouti mouse
3. Microinject ES cells into blastocyst from black-colored mouse
Inner cell mass
4. Transfer into pseudopregnant surrogate mother; birth of chimeras
Chimeras
5. Chimeric mouse bred to black mouse to create mice heterozygous (+/−) for gene knockout
(+/−)
×
(+/−)
(+/+)
×
(+/+)
6. Breed heterozygous mice to produce mice homozygous (−/−) for gene knockout

FIGURE 20.17 A basic strategy for producing a knockout mouse.

The DNA sequence for the gene of interest to be targeted for KO must be known. Scientists also need to know some sequence information about noncoding sequences that flank the gene at its location in the genome. A *targeting vector* is then constructed. The purpose of the targeting vector is to create a segment of DNA that can be introduced into cells. It then undergoes homologous recombination with the gene of interest (the target gene). Recombination disrupts or replaces the gene of interest, thereby rendering it nonfunctional. The targeting vector contains a copy of the gene of interest that has been mutated by inserting a large segment of foreign DNA, essentially a large insertion mutation. This foreign DNA will disrupt the reading frame of the target gene so that if the gene is transcribed into mRNA and translated into protein, it will produce a nonfunctional protein.

To help scientists determine whether or not the targeting vector has been properly introduced into the genome, the insertion sequence typically contains a selectable marker gene. The example shown in Figure 20.17 uses a marker sequence for neomycin resistance (*neo*^{R}). Neomycin is an antibiotic that blocks protein synthesis in both bacterial and eukaryotic cells, and its role will become apparent momentarily. Genes like *GFP*, the gene for green fluorescent protein, the *lacZ* gene discussed earlier in this chapter, among others, are sometimes also used as *reporter genes*. Reporters produce KO or transgenic organisms that are easy to detect visually; you will see an example of a *GFP* transgenic animal later in this section (see Figure 20.20).

There are several ways to introduce the targeting vector into cells. One popular approach involves using electroporation to deliver the vector into cultured **embryonic stem (ES) cells.** The ES cells are harvested from the inner cell mass of a mouse embryo at the blastocyst stage. Alternatively, the targeting vector is directly injected into the blastocyst with the hopes that it will enter ES cells in the inner cell mass. Sometimes it is possible to make KOs by isolating newly fertilized eggs from a female mouse (or female of another desired animal species) and microinjecting the targeting vector DNA directly into the diploid nucleus of the egg or into one of the haploid pronuclei prior to fusion [**Figure 20.18(a)**].

When working with ES cells, the actions of the endogenous enzyme recombinase will catalyze homologous recombination between the targeting vector and the sequence for the gene of interest in only a small percentage of cells that take up the targeting vector. In the few recombinant ES cells that will be created, the targeting vector will usually replace the original gene on only one of two chromosomes.

ES cells can be selected for recombinancy by treating them with a reagent that will kill cells that lack the targeting vector. For the example shown in Figure 20.17, Geneticin™ would be added to cultured ES cells. Recombinant cells containing the targeting vector are resistant to neomycin, but ES cells that are nonrecombinant die. Recombinant ES cells are injected into a mouse embryo at the blastocyst stage where they will be incorporated into the inner cell mass of the blastocyst. The blastocyst is then placed into the uterus of a surrogate mother mouse, sometimes called a *pseudopregnant mouse*—a mouse previously mated with an infertile male to stimulate hormone production in the female mouse that results in a uterus

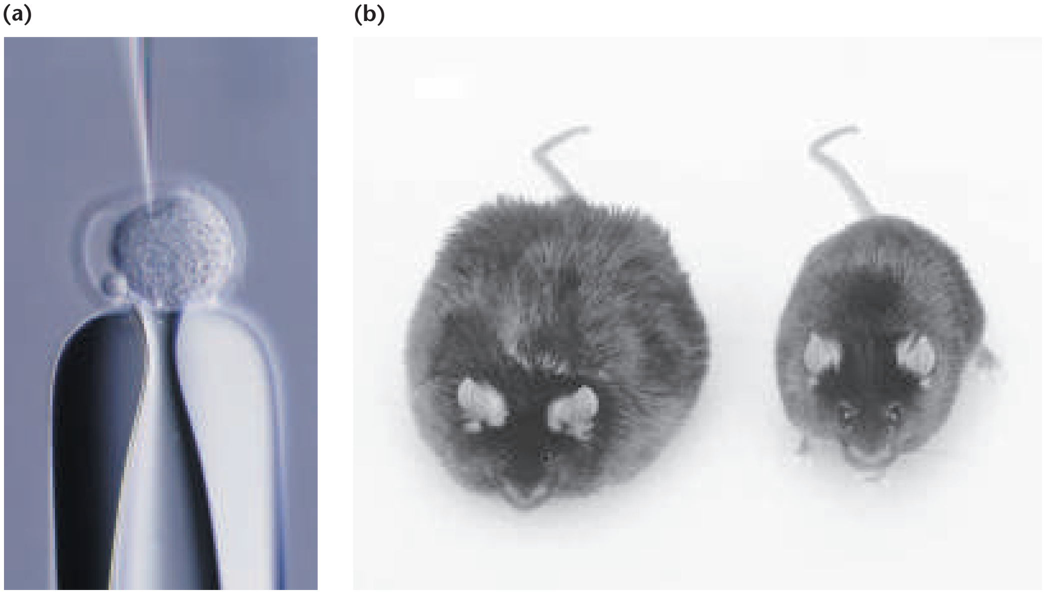


FIGURE 20.18 (a) Microinjecting DNA into a fertilized egg to create a knockout (or a transgenic) mouse. A fertilized egg is held by a suction or holding pipette (seen below the egg), and a microinjection needle delivers cloned DNA into the nucleus. (b) On the left is a null mouse (−/−) for both copies of the obese (*ob*) gene, which produces a peptide hormone called leptin. The mouse on the right is wild type (+/+) for the *Lep* gene. The *Lep* knockout mouse weighs almost five times as much as its wild-type sibling. Naturally occurring mutations in the human *LEP* gene create weight disorders for affected individuals.

receptive to implantation of the blastocyst containing the targeting vector.

The surrogate will give birth to mice that are *chimeras*: Some cells in their body arise from KO stem cells, and others arise from stem cells of the injected blastocyst. As long as the targeting vector DNA is present in germ cells, the sequence will be inherited in all of the offspring generated by these mice, but typically most F_1 generation KO mice produced this way are heterozygous (+/−) for the gene of interest and the targeting vector and not homozygous for the KO. Sibling matings of F_1 animals can then be used to generate homozygous KO animals, referred to as *null mice* and given a −/− designation because they lack wild-type copies of the targeted gene of interest. As mentioned at the beginning of this section, KO animal models serve invaluable roles for learning about gene function, and they continue to be essential for biomedical research on disease genes [see **Figure 20.18(b)**].

Despite all of the work that goes into trying to produce a KO organism, sometimes viable offspring are never born. The KO results in *embryonic lethality*. Knocking out a gene that is important during embryonic development may kill the mouse before researchers have a chance to study it. Typically, researchers will examine embryos from the surrogate mouse to see if they can determine at what stage of embryonic development the embryo is dying. This examination often reveals specific organ defects that can also be informative about the function of the KO gene.

If null mice for a particular gene of interest cannot be derived by traditional KO approaches, an alternative approach called **conditional knockout** can often provide a way to study such a gene. Conditional knockouts allow one to control the particular time in an animal's development that a target gene is disrupted. For example, if a target gene displays embryonic lethality, one can use a conditional KO to allow an animal to progress through development and be born before disrupting. Another advantage of conditional KOs is that target genes can also be turned off in a particular tissue or organ instead of the entire animal.

One common approach for making conditional KOs is called the *Cre-lox system* (**Figure 20.19**). With this method, conditional KOs are made by inserting sequences called *loxP* sites into the targeting vector on either side of the target gene of interest. The gene of interest will be introduced into the germ line via the same mechanism as a knockout. These mice can then be crossed to other mice with a germ line containing the *Cre* gene, which encodes Cre recombinase, a viral enzyme that can recognize the *loxP* sequences, recombines them and deletes the gene between the *loxP* sequences.

A key to the *Cre-lox* system is that the promoter for the *Cre* mouse strain can be designed to be tissue specific (i.e., it will only be expressed in certain tissues). This example is shown in Figure 20.19, where *Cre* is under the control of a tissue-specific promoter X. To prevent embryonic lethality, the promoter chosen for the *Cre* strain can also be under the control of a particular hormone, dietary nutrients, or other conditions. Thus the promoter responsible for expressing *Cre* can be turned on or off at different stages in the development of the animal to produce the KO under desired specific conditions, depending on the research being carried out.

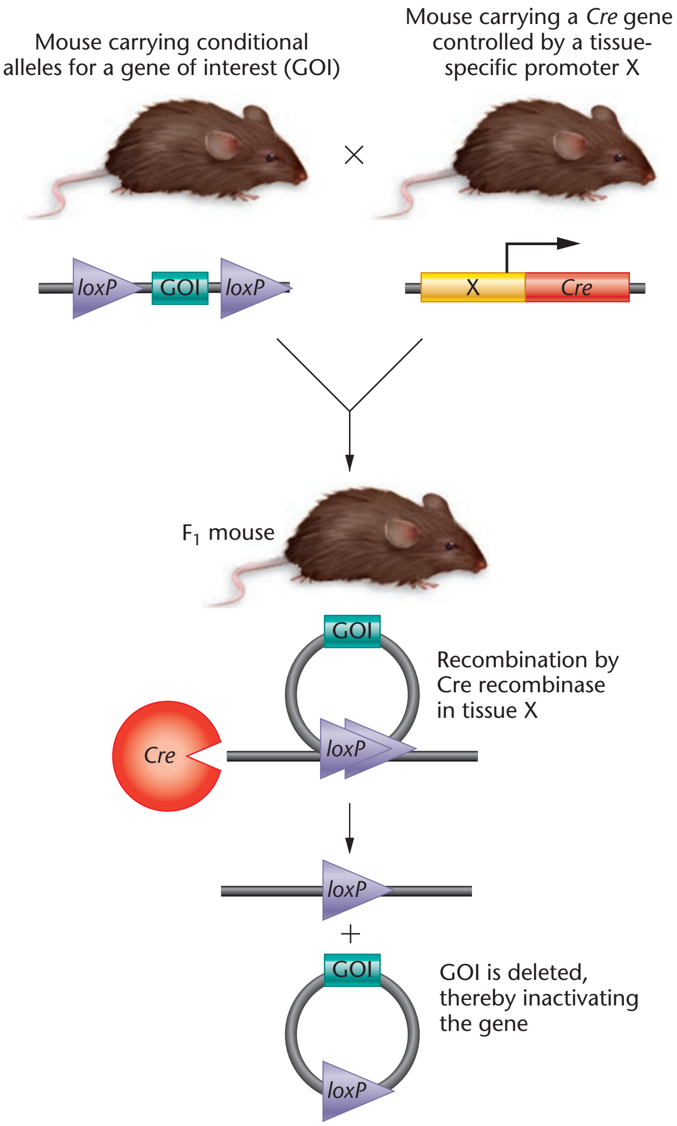

Mouse carrying conditional alleles for a gene of interest (GOI)
Mouse carrying a *Cre* gene controlled by a tissue-specific promoter X
×
loxP
GOI
loxP
X
Cre
F_1 mouse
GOI
Cre
loxP
Recombination by Cre recombinase in tissue X
loxP
+
GOI
loxP
GOI is deleted, thereby inactivating the gene

FIGURE 20.19 The *Cre-lox* system for creating conditional knockouts.

Recently a worldwide collaboration involving more than $100 million in funding to disable all (~20,000) protein-coding genes in the mouse genome was completed. The purpose of the initiative was to create a KO mouse resource that would help advance human disease research. Rather than using a gene-by-gene approach to make KOs, researchers used a high-throughput technique that involves knocking out thousands of genes in embryonic stem cells and then creating offspring in which specific genes of interest can then be turned off. Projects such as this provide a "library" of KO animals as an efficient resource for the research community so that individual scientists can use these animals when possible rather than go through the expense and technical challenges of making their own knockouts.

In addition to creating knockout model organisms for studying disease genes in humans, several research teams are carrying out projects to identify naturally occurring knockouts of human genes. This involves exome sequencing and RNA sequencing (to identify missing non-coding RNAs as well) large populations of individuals who have an inherited disease. Obviously loss-of-function mutations that are lethal cannot be studied in an adult, but some of these projects propose sequencing embryos that did not survive to full term as one approach for identifying lethal gene knockouts. The hope of projects such as these is that they will lead to a catalog or database of phenotypic effects associated with mutations of all human genes.

Making a Transgenic Animal: The Basics

Transgenic animals, also sometimes called **knock-in animals,** express, or often overexpress, a particular gene of interest (the transgene)—in other words, the opposite of KOs, in which a gene's function is turned off. As with KOs, many of the prevailing techniques used to make transgenic animals were developed in mice; **Figure 20.20** illustrates two examples.

The method of creating a transgenic animal is conceptually simple, and many of the steps are similar to the steps involved in making a KO animal. But instead of trying to disrupt a target gene, a vector with the transgene is created to undergo homologous recombination and enter into the host genome. In some applications, tissue-specific promoter sequences can be used so that the transgene is expressed only in specific tissues. For example, in the biotechnology industry tissue-specific promoters are used to express specific recombinant products in milk for subsequent purification. It is often much easier to make a transgenic animal than a KO animal because the vector just needs to be incorporated into the host genome somewhere (hopefully in a noncoding region) but often not at a particular locus as is necessary when making a KO.

The vector with the transgene can be put into ES cells or injected directly into embryos or eggs. As with KOs, the vector will include a marker or reporter gene to help researchers identify successful transformation. Then, in a relatively small percentage of embryos or eggs, the transgenic DNA becomes inserted into the genome by recombination due to the action of naturally occurring DNA recombinases. After this stage, the rest of the process is similar to making a KO: embryos are implanted in surrogate mothers, and crosses from resulting progeny are used to derive mice that are homozygous for the transgene.

In a transgenic experiment, the transgene is often overexpressed in order to study its effects on the appearance and functions of the organism. There are many variations and purposes for making transgenics. Transgenic animals overexpressing certain genes, expressing human genes or genes from a different species, and expressing mutant genes are among examples of transgenics that are valuable models for basic and applied research to understand gene function.

Expressing human genes in transgenic mice is one approach that researchers have used to create *humanized mice* to study responses to different drugs for treating diseases, to understand the roles of genes in evolution, and to study

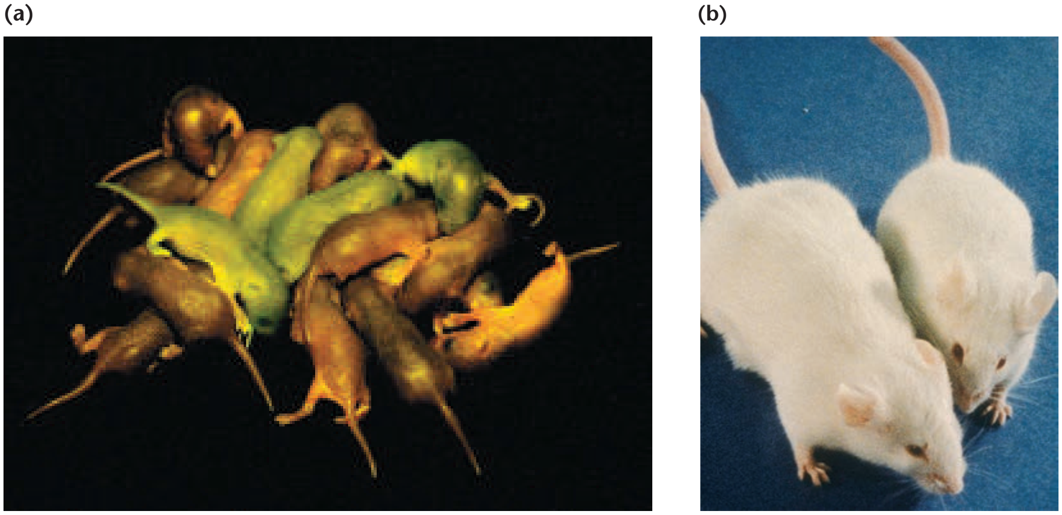


FIGURE 20.20 Examples of transgenic mice. (a) Transgenic mice incorporating the *GFP* gene (a popular reporter gene) from jellyfish enable scientists to tag particular genes with green fluorescent protein. Thanks to the expression of *GFP,* which makes the transgenic mice glow green under ultraviolet light, scientists can track activity of the tagged genes, including activity in subsequent generations of mice generated from these transgenics. (b) The mouse on the left is transgenic for a rat growth hormone gene, cloned downstream from a mouse metallothionein promoter. When the transgenic mouse was fed zinc, the metallothionein promoter induced the transcription of the growth hormone gene, stimulating the growth of the transgenic mouse.

embryonic development, among many other applications. As we will consider later in the text (see Chapter 22), transgenic animals and plants are also created to produce commercially valuable biotechnology products. Also later in the text (see Special Topic Chapter 4—Genetically Modified Foods), you will learn about examples of transgenic food crops.

Gene Editing with CRISPR-Cas

Gene editing methods involve the use of specifically engineered DNA-modifying enzymes (nucleases) that allow researchers to create changes in a specific sequence to remove, correct, or replace a defective gene or parts of a gene. Gene editing is based on using different nucleases to create breaks in the genome in a sequence-specific manner. Later in the text (see Special Topic Chapter 5—Gene Therapy), we discuss how gene editing methods with *transcription activator-like effector nucleases (TALENs)*/ and *zinc finger nucleases (ZFNs)* can be used for gene therapy. These approaches have been used for the *in vivo* genetic engineering of mice, rats, and other species for editing specific genes. But to date, these approaches to gene editing have not been particularly successful for human gene therapy.

No approach to gene editing has garnered more attention, created more excitement, or generated as many early signs of success as the **CRISPR (clustered regularly interspaced palindromic repeats)-Cas system.** In brief, CRISPR-Cas was discovered by scientists trying to understand how bacteria fight viral infection. The original research on this system was not intended to create a gene editing technique. CRISPR-Cas is a bacterial immune response system that provides resistance to foreign genetic material such as plasmids from bacteria and phage DNA.

CRISPRs consist of short base-pair repeats. CRISPR loci contain interspersed spacers (called protospacers) of viral or plasmid DNA, which come from prior exposures to foreign plasmids or phage. Each protospacer is followed by a 2–6 base-pair segment of spacer DNA, called the *protospacer adjacent motif (PAM)* (5′-NGG-3′ where N is any nucleotide). PAM sequences help bacteria distinguish bacterial DNA from foreign DNA and are not part of the CRISPR locus, but are essential sequences for the CRISPR editing process to recognize and remove protospacers.

CRISPR-associated proteins, or Cas proteins, function as nucleases. Cas9 was the first such CRISPR nuclease identified. Cas9 will not bind to or cleave a protospacer DNA sequence if it is not followed by a PAM sequence. In bacteria, CRISPR spacer sequences can recognize viral DNA and use Cas nucleases to digest sequences in the foreign DNA. By delivering the Cas9 nuclease complexed with a synthetic guide RNA (sgRNA) into eukaryotic cells, the eukaryotic genome can likewise be targeted and cut a specific target gene location—thus allowing genes to be removed or new genes to be inserted in cultured cells or in whole organisms (see Figure 20.21). In this manner, scientists are effectively substituting the protospacer for a target gene that they are trying to remove or edit.

The emergence of the CRISPR-Cas system warrants expanded coverage later in the text (see Special Topic Chapter 1—CRISPR-Cas and Genome Editing) including its applications in gene therapy (see Special Topic Chapter 5—Gene Therapy).

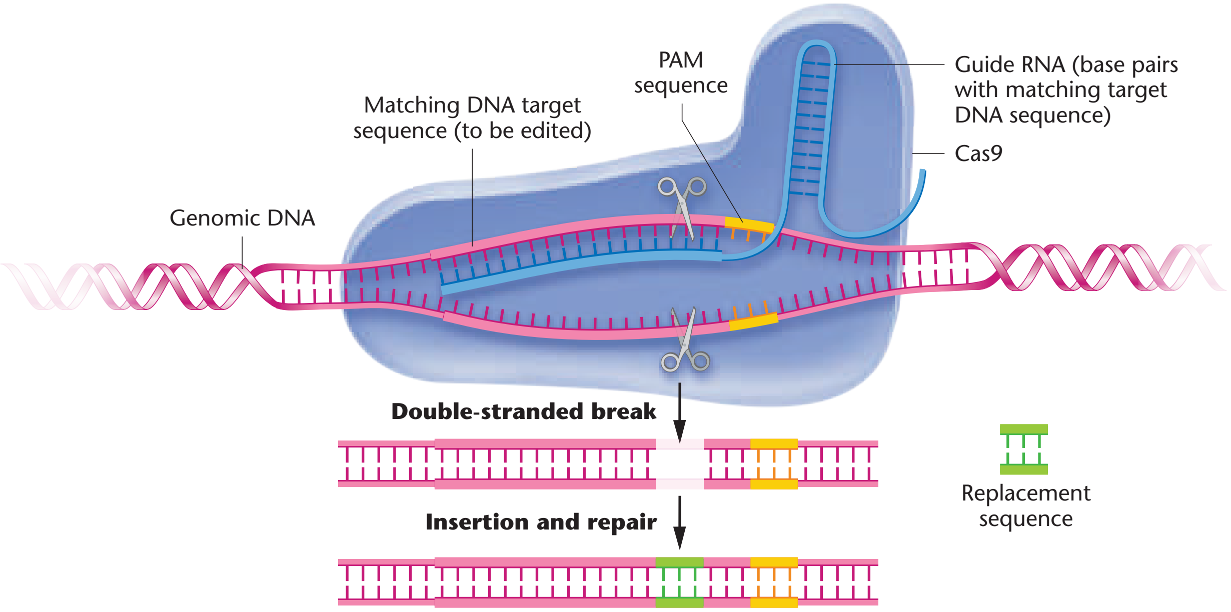

PAM sequence
Guide RNA (base pairs with matching target DNA sequence)
Matching DNA target sequence (to be edited)
Cas9
Genomic DNA
Double-stranded break
Replacement sequence
Insertion and repair

FIGURE 20.21 The CRISPR-Cas9 system allows for gene editing in eukaryotes by targeting specific DNA sequences.

CRISPR-Cas has been a boon for companies providing custom knockouts and transgenic animals to the researcher community. The CRISPR-Cas system is much easier and the success rate of CRISPR-generated knockouts and transgenics is much higher than using ES cells. CRISPR-Cas is also more efficient, reducing the time (~6 months for the CRISPR-Cas approach compared to 18 months or longer to produce a knockout or transgenic animal by traditional ES cell methods) and expense of making these recombinant animals by a third or more. In the future, the emergence of CRISPR-Cas may render current approaches for making both knockout and transgenic animals obsolete.

GNNANACTGACNCAC TA TAGGGCGAATTCGAGCTCGGTACCCGGNGGATCCTCTAGAG TCGACC
10
20
30
40
50
60
70
80

EXPLORING GENOMICS

Mastering Genetics Visit the Study Area: Exploring Genomics

Manipulating Recombinant DNA: Restriction Mapping and Designing PCR Primers

As you learned in this chapter, restriction enzymes are sophisticated "scissors" that molecular biologists use to cut DNA, and they are routinely used in genetics and molecular biology laboratories for recombinant DNA experiments. A wide variety of online tools assist scientists working with restriction enzymes and manipulating recombinant DNA for different applications. Here we explore **Webcutter** and **Primer3**, two sites that make recombinant DNA experiments much easier.

Exercise I – Creating a Restriction Map in Webcutter

Suppose you had cloned and sequenced a gene and you wanted to design a probe approximately 600 bp long that could be used to analyze expression of this gene in different human tissues by Northern blot analysis. Internet sites such as Webcutter make it relatively easy to design experiments for manipulating recombinant DNA. In this exercise, you will use Webcutter to create a restriction map of human DNA with the enzymes *Eco*RI, *Bam*HI, and *Pst*I.

1. Access **Webcutter** at http://rna.lundberg.gu.se/cutter2/. Go to the Study Area for *Concepts of Genetics,* and open the Exploring Genomics exercise for this chapter. Copy the sequence of cloned human DNA found there, and paste it into the text box in Webcutter.
2. Scroll down to "Please indicate which enzymes to include in the analysis." Click the button indicating "Use only the following enzymes." Select the restriction enzymes *Eco*RI, *Bam*HI, and *Pst*I from the list provided, and then click "Analyze sequence." (*Note*: Use the command, control, or shift key to select multiple restriction enzymes.)
3. After examining the results provided by Webcutter, create a table showing the number of cutting sites for each enzyme and the fragment sizes that would be generated by digesting with each enzyme. Draw a restriction map indicating cutting sites for each enzyme with distances between each site and the total size of this piece of human DNA.

Exercise II – Designing a Recombinant DNA Experiment

Now that you have created a restriction map of your piece of human DNA, you need to ligate the DNA into a plasmid DNA vector that you can use to make your probe (molecular biologists often refer to this as subcloning). To do this, you will need to determine which restriction enzymes would best be suited for cutting both the plasmid and the human DNA.

1. Referring back to the Study Area and the Exploring Genomics exercise for this chapter, copy the plasmid DNA sequence from Exercise I into the text box in Webcutter and identify cutting sites for the same enzymes you used in Exercise I. Then answer the following questions:
   a. What is the total size of the plasmid DNA analyzed in Webcutter?
   b. Which enzyme(s) could be used in a recombinant DNA experiment to ligate the plasmid to the *largest* DNA fragment from the human gene? Briefly explain your answer.
   c. What size recombinant DNA molecule will be created by ligating these fragments?
   d. Draw a simple diagram showing the cloned DNA inserted into the plasmid, and indicate the restriction-enzyme cutting site(s) used to create this recombinant plasmid.

Exercise III – Designing PCR Primers

Suppose you decided to try reverse transcription PCR (RT-PCR) first instead of Northern blotting because RT-PCR is

(continued)

Exploring Genomics—continued

a faster and more sensitive way to detect gene expression. You will now need to design primers for RT-PCR. Picking correct primers for any PCR experiment is not a trivial process. You have to be sure the primers can amplify the gene of interest, and you need to avoid primer self-annealing—or having primers bind to each other—among many other considerations. Fortunately, primer design is another task made much easier by the Internet. In this exercise, you will use Primer3, a PCR primer design site from the Whitehead Institute for Biomedical Research.

1. Access **Primer3** at http://bioinfo.ut.ee/primer3/. Copy the human DNA sequence from Exercise I into the text box, and then click "Pick Primers."
2. On the next page, the sequences for the best recommended primers will appear at the top of the screen. Answer the following:
   a. What is the length, in base pairs, of the left (forward) primer and right (reverse) primer? Where does each of these primers bind in the gene sequence?
   b. The hybridization temperature for a PCR reaction is often set around 5 degrees below the melting temperature, or T_m (refer to Chapter 10 for a discussion of melting temperature). Based on the T_m for these primers, what might be the optimal hybridization temperature for this experiment?
   c. What size PCR product would you expect these primers to generate if you ran the DNA amplified by this PCR reaction on an agarose gel?

CASE STUDY Ethical issues and genetic technology

In the 1970s, scientists realized that there may be unforeseen dangers and ethical issues with the use of recombinant DNA technology. A self-imposed moratorium on related research was implemented to develop safety protocols. As the Human Genome Project, designed to sequence and analyze the DNA of the human genome, came into existence in 1990, it was accompanied by the Ethical, Legal, and Social Implications (ELSI) program. ELSI was charged with identifying and addressing issues arising from genomic research.

This program focused mainly on privacy issues, the ethical use of genetic technology in medicine, and the design and conduct of genetic research, including gene therapy. The program led to the passage of federal legislation regulating the use of genetic information, and instituting guidelines limiting the scope of gene therapy. These guidelines prohibit germ-line therapy, which impact future generations, and also prohibit gene therapy designed to enhance physical or mental aptitudes.

However, recently, the development of CRISPR-Cas as a new genetic technology may allow for the removal of mutant alleles that cause devastating neurological disorders such as Huntington disease and prevent its transmission to future generations. Similar technology can be used to selectively eradicate the species of mosquito that transmits malaria, a painful and life-shortening disease that affects millions worldwide. With the development of these revolutionary methods, there are new calls to redefine and institute a new set of ethical guidelines for using these methods to eliminate genetic disorders and to revolutionize agriculture.

1. What undesirable or unforeseen consequences might occur in ecosystems if a species is eradicated using these new technologies?
2. Do we have the ethical right to alter the genomes of future generations of humans even if intervention eliminates lethal alleles?
3. Should these new technologies be regulated internationally to prevent their use by bioterrorists? How could violations be detected, and how could such regulations be enforced?

For related reading, see Rodriguez, E. (2016). Ethical issues in using CRISPR/Cas9 system. *J. Clin. Res. Bioethics* 7:266 (doi:10.4172/2155-9627.1000266).

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Recombinant DNA technology was made possible by the discovery of specific proteins called restriction enzymes, which cut DNA at specific recognition sequences, producing fragments that can be joined with other DNA fragments to form recombinant DNA molecules.
2. Recombinant DNA molecules can be transferred into any of several types of host cells where cloned copies of the DNA are produced during host-cell replication. Many kinds of host cells may be used for replication, including bacteria, yeast, and mammalian cells.
3. Historically, DNA libraries have been important for producing collections of cloned genes to identify genes and gene-regulatory regions of interest. Aspects of genomic analysis still rely on applications of DNA libraries.
4. The polymerase chain reaction (PCR) allows DNA to be amplified without host cells and is a rapid, sensitive method with wide-ranging applications.
5. Once cloned, DNA sequences are analyzed through a variety of molecular techniques that allow scientists to study gene structure, expression, and function.
6. By determining the nucleotide sequence of a DNA segment, DNA sequencing is the ultimate way to characterize DNA at the molecular level.
7. Rapid advances in sequencing technologies have led to greatly increased sequencing capacities at reduced costs over historically used sequencing methods, providing scientists with unprecedented access to sequence data.
8. Gene knockout methods and transgenic animals are widely used to study gene function *in vivo*.
9. Gene editing approaches have emerging roles in basic research and in clinical applications of genetic technology.

INSIGHTS AND SOLUTIONS

The recognition sequence for the restriction enzyme *Sau*3AI is GATC (see Figure 20.1); in the recognition sequence for the enzyme *Bam*HI—GGATCC—the four internal bases are identical to the *Sau*3AI sequence. The single-stranded ends produced by the two enzymes are identical. Suppose you have a cloning vector that contains a *Bam*HI recognition sequence and you also have foreign DNA that was cut with *Sau*3AI.

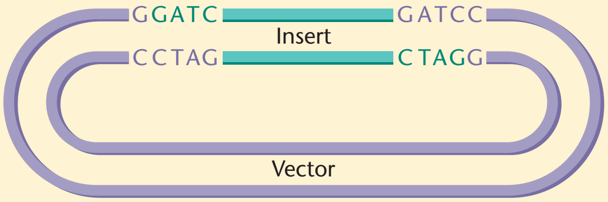
(a) Can this DNA be ligated into the *Bam*HI site of the vector, and if so, why?

(b) Can the DNA segment cloned into this sequence be cut from the vector with *Sau*3AI? With *Bam*HI? What potential problems do you see with the use of *Bam*HI?

Solution:

(a) DNA cut with *Sau*3AI can be ligated into the vector's *Bam*HI cutting site because the single-stranded ends generated by the two enzymes are identical.

(b) The DNA can be cut from the vector with *Sau*3AI because the recognition sequence for this enzyme (GATC) is maintained on each side of the insert. Cutting the cloned insert with *Bam*HI is more problematic. In the ligated vector, the conserved sequences are GGATC (left) and GATCC (right). The correct base for recognition by *Bam*HI will *follow* the conserved sequence (to produce GGATCC on the left) only about 25 percent of the time, and the correct base will *precede* the conserved sequence (and produce GGATCC on the right) about 25 percent of the time as well. Thus, *Bam*HI will be able to cut the insert from the vector (0.25 × 0.25 = 0.0625), or only about 6 percent, of the time.



Problems and Discussion Questions

1. HOW DO WE KNOW? In this chapter we focused on how specific DNA sequences can be copied, identified, characterized, and sequenced. At the same time, we found many opportunities to consider the methods and reasoning underlying these techniques. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) In a recombinant DNA cloning experiment, how can we determine whether DNA fragments of interest have been incorporated into plasmids and, once host cells are transformed, which cells contain recombinant DNA?
(b) What steps make PCR a chain reaction that can produce millions of copies of a specific DNA molecule in a matter of hours without using host cells?
(c) How has DNA-sequencing technology evolved in response to the emerging needs of genome scientists?
(d) How can gene knockouts, transgenic animals, and gene editing techniques be used to explore gene function?

2. CONCEPT QUESTION Review the Chapter Concepts list on page 454. All of these refer to recombinant DNA methods and applications. Write a short essay or sketch a diagram that provides an overview of how recombinant DNA techniques help geneticists study genes.

3. What roles do restriction enzymes, vectors, and host cells play in recombinant DNA studies? What role does DNA ligase perform in a DNA cloning experiment? How does the action of DNA ligase differ from the function of restriction enzymes?

4. The human insulin gene contains a number of sequences that are removed in the processing of the mRNA transcript. In spite of the fact that bacterial cells cannot excise these sequences from mRNA transcripts, explain how a gene like this can be cloned into a bacterial cell and produce insulin.

5. Although many cloning applications involve introducing recombinant DNA into bacterial host cells, many other cell types are also used as hosts for recombinant DNA. Why?

6. Using DNA sequencing on a cloned DNA segment, you recover the nucleotide sequence shown below. Does this segment contain a palindromic recognition sequence for a restriction enzyme? If so, what is the double-stranded sequence of the palindrome, and what enzyme would cut at this sequence? (Consult Figure 20.1 for a list of restriction sites.)

CAGTATGGATCCCAT

7. Restriction sites are palindromic; that is, they read the same in the 5′ to 3′ direction on each strand of DNA. What is the advantage of having restriction sites organized this way?

8. List the advantages and disadvantages of using plasmids as cloning vectors. What advantages do BACs and YACs provide over plasmids as cloning vectors?

9. What are the advantages of using a restriction enzyme whose recognition site is relatively rare? When would you use such enzymes?

10. In 1975, the Asilomar Conference on Recombinant DNA was organized by Paul Berg, a pioneer of recombinant DNA technology, at a conference center at Asilomar State Beach in California. Physicians, scientists, lawyers, ethicists, and others gathered to draft guidelines for safe applications of recombinant DNA technology. These general guidelines were adopted by the federal government and are still in practice today. Consider the implications of recombinant DNA as a new technology. What concerns might the scientific community have had then about recombinant DNA technology? Might those same concerns exist today?

11. In the context of recombinant DNA technology, of what use is a probe?

12. If you performed a PCR experiment starting with only one copy of double-stranded DNA, approximately how many DNA molecules would be present in the reaction tube after 15 cycles of amplification?

13. In a control experiment, a plasmid containing a *Hin*dIII recognition sequence within a kanamycin resistance gene is cut with *Hin*dIII, re-ligated, and used to transform *E. coli* K12 cells. Kanamycin-resistant colonies are selected, and plasmid DNA from these colonies is subjected to electrophoresis. Most of the colonies contain plasmids that produce single bands that migrate at the same rate as the original intact plasmid. A few colonies, however, produce two bands, one of original size and one that migrates much less far down the gel. Diagram the origin of this slow band as a product of ligation.

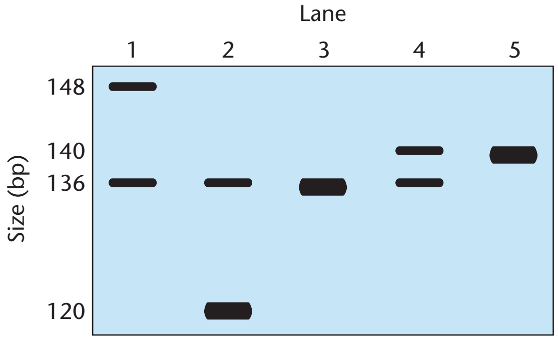
14. What advantages do cDNA libraries provide over genomic DNA libraries? Describe cloning applications where the use of a genomic library is necessary to provide information that a cDNA library cannot.

15. You have recovered a cloned DNA segment from a vector and determine that the insert is 1300 bp in length. To characterize this cloned segment, you isolate the insert and decide to construct a restriction map. Using enzyme I and enzyme II, followed by gel electrophoresis, you determine the number and size of the fragments produced by enzymes I and II alone and in combination, as recorded in the following table. Construct a restriction map from these data, showing the positions of the restriction-enzyme cutting sites relative to one another and the distance between them in units of base pairs.

Enzyme	Restriction Fragment Sizes (bp)
I	350, 950
II	200, 1100
I and II	150, 200, 950

16. To create a cDNA library, cDNA can be inserted into vectors and cloned. In the analysis of cDNA clones, it is often difficult to find clones that are full length—that is, many clones are shorter than the mature mRNA molecules from which they are derived. Why is this so?

17. Although the capture and trading of great apes has been banned in 112 countries since 1973, it is estimated that about 1000 chimpanzees are removed annually from Africa and smuggled into Europe, the United States, and Japan. This illegal trade is often disguised by simulating births in captivity. Until recently, genetic identity tests to uncover these illegal activities were not used because of the lack of highly polymorphic markers (markers that vary from one individual to the next) and the difficulties of obtaining chimpanzee blood samples. A study was reported in which DNA samples were extracted from freshly plucked chimpanzee hair roots and used as templates for PCR. The primers used in these studies flank highly polymorphic sites in human DNA that result from variable numbers of tandem nucleotide repeats. Several offspring and their putative parents were tested to determine whether the offspring were "legitimate" or the product of illegal trading. The data are shown in the following Southern blot.



Lane 1: father chimpanzee
Lane 2: mother chimpanzee
Lanes 3–5: putative offspring A, B, C

Examine the data carefully and choose the best conclusion.

(a) None of the offspring are legitimate.

(b) Offspring B and C are not the products of these parents and were probably purchased on the illegal market. The data are consistent with offspring A being legitimate.

(c) Offspring A and B are products of the parents shown, but C is not and was therefore probably purchased on the illegal market.

(d) There are not enough data to draw any conclusions. Additional polymorphic sites should be examined.

(e) No conclusion can be drawn because "human" primers were used.

18. To estimate the number of cleavage sites in a particular piece of DNA with a known size, you can apply the formula $N/4^n$ where N is the number of base pairs in the target DNA and n is the number of bases in the recognition sequence of the restriction enzyme. If the recognition sequence for *Bam*HI is GGATCC and the λ phage DNA contains approximately 48,500 bp, how many cleavage sites would you expect?

19. In a typical PCR reaction, describe what is happening in stages occurring at temperature ranges (a) 92–95°C, (b) 45–65°C, and (c) 65–75°C.

20. We usually think of enzymes as being most active at around 37°C, yet in PCR the DNA polymerase is subjected to multiple exposures of relatively high temperatures and seems to function appropriately at 65–75°C. What is special about the DNA polymerase typically used in PCR?

21. Traditional Sanger sequencing has largely been replaced in recent years by next-generation and third-generation sequencing approaches. Describe advantages of these sequencing methods over first-generation Sanger sequencing.

22. How is fluorescent *in situ* hybridization (FISH) used to produce a spectral karyotype?

23. What is the difference between a knockout animal and a transgenic animal?

24. One complication of making a transgenic animal is that the transgene may integrate at random into the coding region, or the regulatory region, of an endogenous gene. What might be the consequences of such random integrations? How might this complicate genetic analysis of the transgene?

25. When disrupting a mouse gene by knockout, why is it desirable to breed mice until offspring homozygous (−/−) for the knockout target gene are obtained?

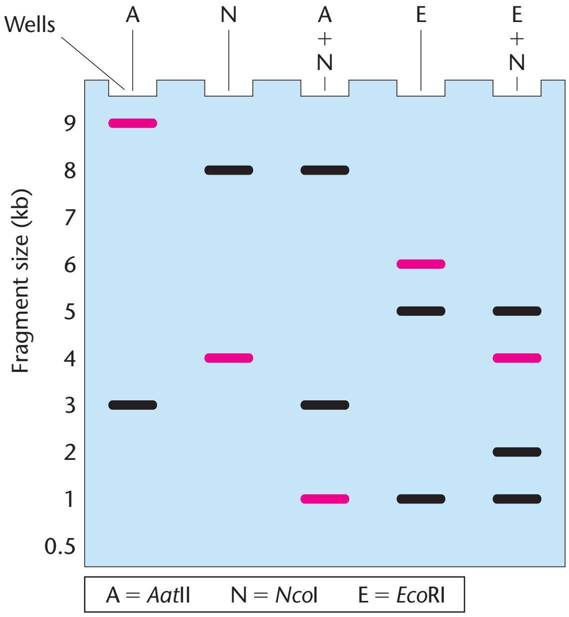
26. What techniques can scientists use to determine if a particular transgene has been integrated into the genome of an organism?

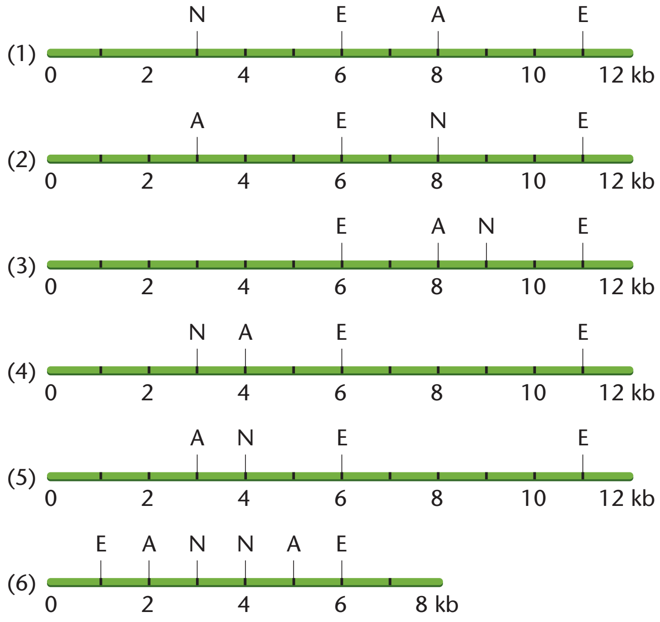
27. Gene targeting and gene editing are both techniques for removing or modifying a particular gene, each of which can produce the same ultimate goal. What is the main technical difference in how DNA is modified that differs between these approaches?

28. As you will learn later in the text (Special Topics Chapter 1—CRISPR-Cas and Genome Editing), the CRISPR-Cas system has great potential but also raises many ethical issues about its potential applications because theoretically it can be used to edit any gene in the genome. What do you think are some of the concerns about the use of CRISPR-Cas on humans? Should CRISPR-Cas applications be limited for use on only certain human genes but not others? Explain your answers.

Extra-Spicy Problems

29. The gel presented here shows the pattern of bands of fragments produced with several restriction enzymes. The enzymes used are identified above the lanes of the gel, and six possible restriction maps are shown in the column to the right.





One of the six restriction maps shown is consistent with the pattern of bands shown in the gel.

(a) From your analysis of the pattern of bands on the gel, select the correct restriction map and explain your reasoning.

(b) The highlighted bands (magenta) in the gel hybridized with a probe for the gene *pep* during a Southern blot. Where in the gel is the *pep* gene located?

30. A widely used method for calculating the annealing temperature for a primer used in PCR is 5 degrees below the melting temperature, T_m (°C), which is computed by the equation $81.5 + 0.41 \times (\%GC) - (675/N)$, where %GC is the percentage of GC nucleotides in the oligonucleotide and N is the length of the oligonucleotide. Notice from the formula that both the GC content *and* the length of the oligonucleotide are variables. Assuming you have the following oligonucleotide as a primer,

5′-TTGAAAATATTTCCCATTGCC-3′

compute the annealing temperature for PCR. What is the relationship between T_m (°C) and %GC? Why? (*Note:* In reality, this computation provides only a starting point for empirical determination of the most useful annealing temperature.)

31. Most of the techniques described in this chapter (blotting, cloning, PCR, etc.) are dependent on hybridization (annealing) between different populations of nucleic acids. Length of the strands, temperature, and percentage of GC nucleotides weigh considerably on hybridization. Two other components commonly used in hybridization protocols are monovalent ions and formamide. A formula that takes monovalent Na^+ ions ((M[Na^+]) and formamide concentrations into consideration to compute a T_m (temperature of melting) is as follows:

$$T_m = 81.5 + 16.6(\log M[Na^+]) + 0.41(\%GC) - 0.72(\%formamide)$$

(a) For the following concentrations of Na^+ and formamide, calculate the T_m. Assume 45% GC content.

[Na^+]	% Formamide
0.825	20
0.825	40
0.165	20
0.165	40

(b) Given that formamide competes for hydrogen bond locations on nucleic acid bases and monovalent cations are attracted to the negative charges on nucleic acids, explain why the T_m varies as described in part (a).

32. In humans, congenital heart disease is a common birth defect that affects approximately 1 out of 125 live births. Using reverse transcription PCR (RT-PCR) Samir Zaidi and colleagues [(2013) *Nature* 498:220.223] determined that approximately 10 percent of the cases resulted from point mutations, often involving histone function. To capture products of gene expression in developing hearts, they used oligo(dT) in their reverse transcription protocol.

(a) How would such a high %T in a primer influence annealing temperature?

(b) Compared with oligo(dT) primers, a pool of random sequence primers requires a trickier assessment of annealing temperature. Why?

(c) If one were interested in comparing the quantitative distribution of gene expression in say, the right and left side of a developing heart, how might one proceed using RT-PCR?

33. The U.S. Department of Justice has established a database that catalogs PCR amplification products from short tandem repeats of the Y chromosome (Y-STRs) in humans. The database contains polymorphisms of five U.S. ethnic groups (African-Americans, European Americans, Hispanics, Native Americans, and Asian-Americans) as well as the worldwide population.

(a) Given that STRs are repeats of varying lengths, for example $(TCTG)_{9-17}$ or $(TAT)_{6-14}$, explain how PCR could reveal differences (polymorphisms) among individuals. How could the Department of Justice make use of those differences?

(b) Y-STRs from the nonrecombining region of the Y chromosome (NRY) have special relevance for forensic purposes. Why?

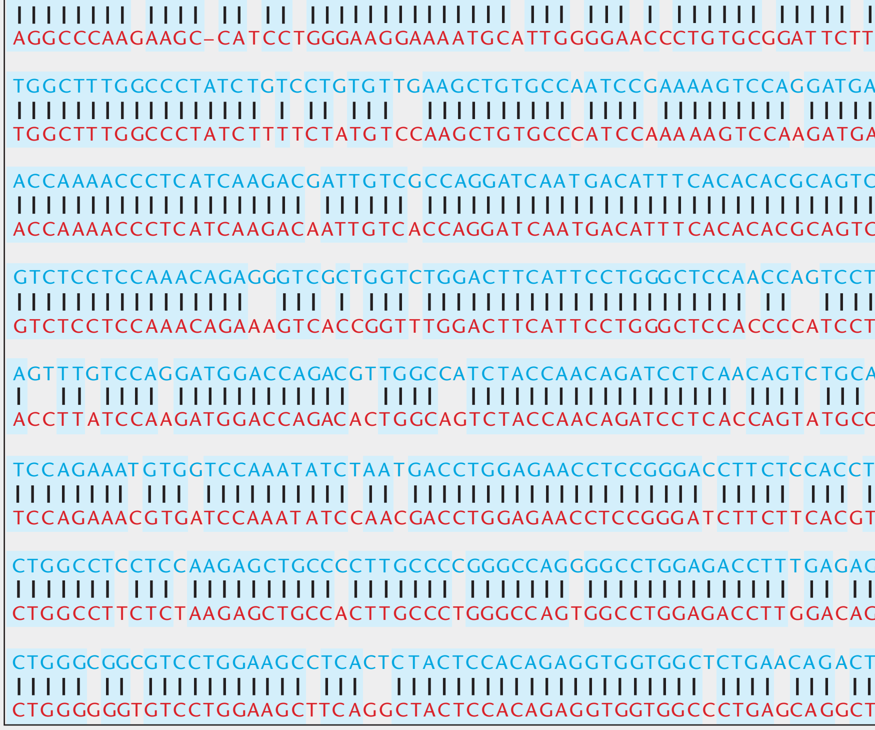
(c) What would be the value of knowing the ethnic population differences for Y-STR polymorphisms?

(d) For forensic applications, the probability of a "match" for a crime scene DNA sample and a suspect's DNA often culminates in a guilty or innocent verdict. How is a "match" determined, and what are the uses and limitations of such probabilities?

34. There are a variety of circumstances under which rapid results using multiple markers in PCR amplifications are highly desired, such as in forensics, pathogen analysis, or detection of genetically modified organisms. In multiplex PCR, multiple sets of primers are used, often with less success than when applied to PCR as individual sets. Numerous studies have been conducted to optimize procedures, but each has described the process as time consuming and often unsuccessful. Considering the information given in Problem 30, why should multiplex PCR be any different than single primer set PCR in terms of dependability and ease of optimization?

21

Genomic Analysis


AGGCCCAAGAAGC–CATCCTGGGAAGGAAAATGCATTGGGGAACCCTGTGCGGATTCTT
TGGCTTTGGCCCTATCTGTCCTGTGTTGAAGCTGTGCCAATCCGAAAAGTCCAGGATGA
TGGCTTTGGCCCTATCTTTTCTATGTCCAAGCTGTGCCCATCCAAAAAGTCCAAGATGA
ACCAAAACCCTCATCAAGACGATTGTCGCCAGGATCAATGACATTTCACACACGCAGTC
ACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTCACACACGCAGTC
GTCTCCTCCAAACAGAGGGTCGCTGGTCTGGACTTCATTCCTGGGCTCCAACCAGTCCT
GTCTCCTCCAAACAGAAAGTCACCGGTTTGGACTTCATTCCTGGGCTCCACCCCATCCT
AGTTTGTCCAGGATGGACCAGACGTTGGCCATCTACCAACAGATCCTCAACAGTCTGCA
ACCTTATCCAAGATGGACCAGACACTGGCAGTCTACCAACAGATCCTCACCAGTATGCC
TCCAGAAATGTGGTCCAAATATCTAATGACCTGGAGAACCTCCGGGACCTTCTCCACCT
TCCAGAAACGTGATCCAAATATCCAACGACCTGGAGAACCTCCGGGATCTTCTTCACGT
CTGGCCTCCTCCAAGAGCTGCCCCTTGCCCCGGGCCAGGGGCCTGGAGACCTTTGAGAC
CTGGCCTTCTCTAAGAGCTGCCACTTGCCCTGGGCCAGTGGCCTGGAGACCTTGGACAC
CTGGGCGGCGTCCTGGAAGCCTCACTCTACTCCACAGAGGTGGTGGCTCTGAACAGACT
CTGGGGGGTGTCCTGGAAGCTTCAGGCTACTCCACAGAGGTGGTGGCCCTGAGCAGGCT

Alignment comparing the DNA sequence for the leptin gene from dogs (blue) and from humans (red). Vertical lines and shaded boxes indicate identical bases. *LEP* encodes a hormone that functions to suppress appetite. This type of analysis is a common application of bioinformatics and a good demonstration of comparative genomics.

CHAPTER CONCEPTS

- Genomics applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze genomes.
- Disciplines in genomics encompass several areas of study, including structural and functional genomics, comparative genomics, and metagenomics, and have led to an "omics" revolution in modern biology.
- Bioinformatics merges information technology with biology and mathematics to store, share, compare, and analyze nucleic acid and protein sequence data.
- The Human Genome Project has greatly advanced our understanding of the organization, size, and function of the human genome.
- Fifteen years after completion of the Human Genome Project, a new era of genomics studies is providing deeper insights into the human genome.
- Comparative genomics analysis has revealed similarities and differences in genome size and organization.
- Metagenomics is the study of genomes from environmental samples and is valuable for identifying microbial genomes.
- Transcriptome analysis provides insight into patterns of gene expression and gene-regulatory activity of a genome.
- Proteomics focuses on the protein content of cells and on the structures, functions, and interactions of proteins.

In 1977, as recombinant DNA–based techniques were developing, including DNA sequencing, Fred Sanger and colleagues launched the field of **genomic analysis** or simply **genomics,** the study of genomes, by sequencing the 5400-nucleotide DNA genome of the virus ϕX174. Over the past 40 years, genomic analysis has advanced so quickly that modern biological research is experiencing a genomics revolution. Genomics is one of the most rapidly advancing and exciting areas of modern genetics—providing scientists, and even the general public, with unprecedented information about genomes of different organisms, including humans, on a daily basis.

In this chapter, we will examine basic technologies used in genomic analysis, discuss examples of genome data derived from different species including the human genome, and consider selected disciplines of genomics. We conclude by discussing *transcriptome analysis*, the study of genes expressed in a cell or tissue (the "transcriptome"), and *proteomics*, the study of proteins present in a cell or tissue.

We will continue our discussion of genomics (in Chapter 22) by highlighting modern applications of recombinant DNA and genomics. Note that some of the topics discussed in this chapter are explored in greater depth in other chapters (see Chapter 19 and Special Topic Chapter 3—Genomics and Precision Medicine).

21.1 Genomic Analysis Before Modern Sequencing Methods Involved Classical Genetics Approaches and Cloning to Map One or a Few Genes at a Time

Prior to the development of early methods for DNA sequencing (such as Sanger sequencing), geneticists typically followed a two-part classical genetics approach to identify and characterize all of the genes in an organism's genome: (1) they identified spontaneous mutations or collected mutants induced by chemical or physical agents, and (2) they generated linkage maps using these mutant strains (as discussed in Chapter 5).

Such strategies were used to identify genes in many model organisms, such as *Drosophila*, maize, mice, bacteria, and yeast, as well as in viruses such as bacteriophages. These approaches were essential for analyzing genomes; however, they had several major limitations. For instance, mutational analysis and linkage requires that at least one mutation for each gene be available before additional genes in a genome can be identified. Obtaining mutants and carrying out linkage studies are very time consuming processes, and when mutations are lethal or have no clear phenotype, they can be difficult or impossible to map.

In addition, although researchers can generate mutations in animal models in a laboratory, they cannot do the same with humans; thus, identifying human genes by mutational analysis is largely limited to linkage mapping of inherited or spontaneously acquired mutant genes with clear phenotypes. Another fundamental limitation of the classical genetics approaches is that they do not lead to a determination of DNA sequence. Nor are they particularly useful for studying noncoding areas of the genome such as DNA regulatory sequences.

In the 1980s, geneticists interested in mapping human genes began using recombinant DNA technology to map DNA sequences to specific chromosomes. Initially, most of these sequences were not actually full-length genes but marker sequences such as *restriction fragment length polymorphisms (RFLPs)*. Once assigned to chromosomes, these markers were used in pedigree analysis to establish linkages between the markers and disease phenotypes for genetic disorders. This approach, called *positional cloning*, was used to map, isolate, clone, and sequence the genes for cystic fibrosis, neurofibromatosis, and dozens of other disorders. Positional cloning identified one gene at a time, and yet by the mid-1980s, it had been used to assign more than 3500 genes and markers to human chromosomes. Recombinant DNA technology also made it possible to generate DNA libraries that could be used to identify, clone, and sequence specific genes of interest. In the 1990s it was estimated that there were approximately 100,000 genes in the human genome which was later found to be inaccurate, and it was readily apparent that mapping and cloning the human genome using existing methods would be a laborious, time-consuming, and nearly insurmountable task. But rapid advances in DNA-sequencing methods made it possible to consider sequencing the larger and more complex genomes of eukaryotes, including the human genome. The development of new DNA-sequencing methods and bioinformatics is responsible for modern genomic analysis.

21.2 Whole-Genome Sequencing Is Widely Used for Sequencing and Assembling Entire Genomes

A primary limitation of most recombinant DNA approaches is that they typically can identify only relatively small numbers of genes at a time. Genomics strategies allow the sequencing of entire genomes. The most widely used strategy for sequencing and assembling an entire genome involves variations of a method called **whole-genome sequencing (WGS),** also known as shotgun cloning or shotgun sequencing. In simple terms, this technique is analogous to you and a friend taking your respective copies of this genetics textbook and randomly ripping the pages into strips about 5 to 7 inches long. Each chapter represents a chromosome, and all the letters in the entire book are the "genome." Then you and your friend would go through the painstaking task of comparing the pieces of paper to find places that match, overlapping sentences—areas where there are similar sentences on different pieces of paper. Eventually, in theory, many of the strips containing matching sentences would overlap in ways that you could use to reconstruct the pages and assemble the order of the entire text.

Figure 21.1 shows a basic overview of WGS. First, multiple copies of an entire chromosome are cut into short, overlapping fragments, by mechanically shearing the DNA in various ways. For simplicity, here we present an example using restriction enzymes. (Nonenzymatic approaches for shearing DNA, such as excessive heat treatment or sonication in which sonic energy is used to break DNA, are also used.) Different restriction enzymes can be used so that chromosomes are cut at different sites, or sometimes, *partial digests* of DNA using the same restriction enzyme are used. With partial digests, DNA is incubated for only a short period of time so that not every target site in a particular sequence is cut to completion by the restriction enzyme. Restriction digests of whole chromosomes generate thousands to millions of overlapping DNA fragments.

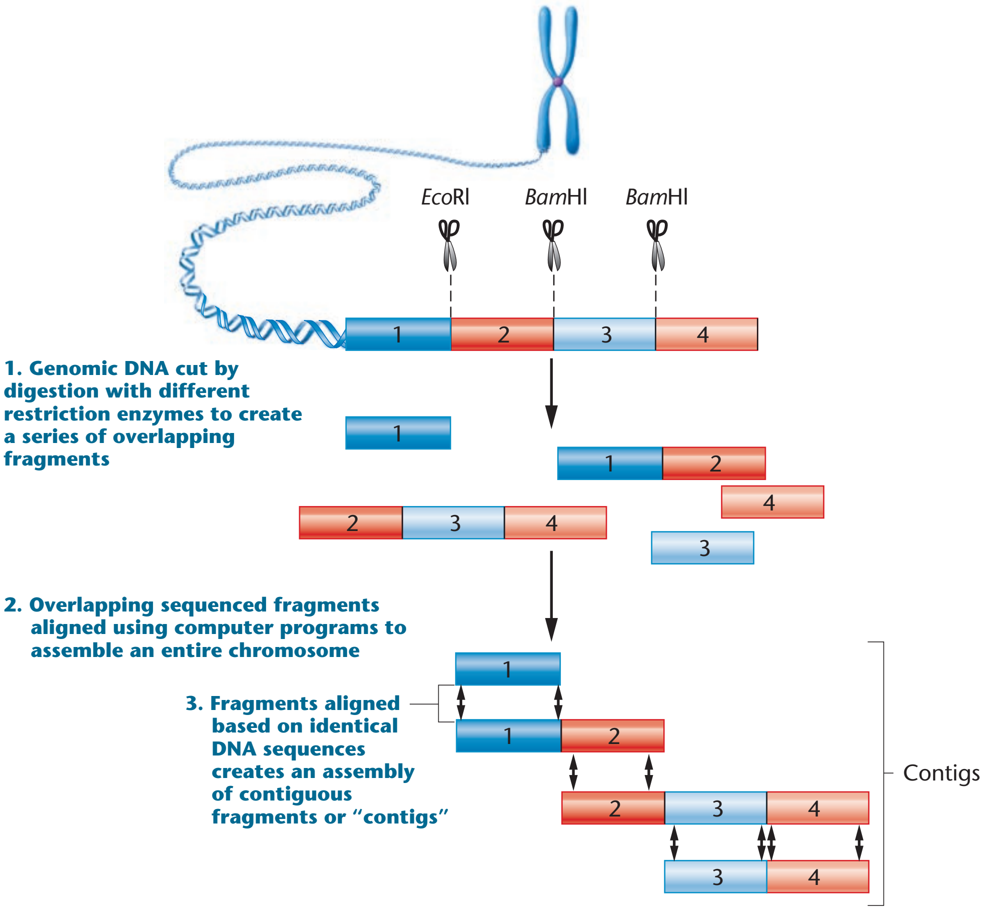

EcoRI
BamHI
BamHI
1
2
3
4
1. Genomic DNA cut by digestion with different restriction enzymes to create a series of overlapping fragments
2. Overlapping sequenced fragments aligned using computer programs to assemble an entire chromosome
3. Fragments aligned based on identical DNA sequences creates an assembly of contiguous fragments or "contigs"
Contigs

FIGURE 21.1 An overview of whole-genome sequencing and assembly. One strategy (shown here) involves using restriction enzymes to digest genomic DNA into contigs, which are then sequenced and aligned using bioinformatics to identify overlapping fragments based on sequence identity. Notice that *Eco*RI digestion produces two fragments (contigs 1 and 2–4), whereas digestion with *Bam*HI produces three fragments (contigs 1–2, 3, and 4).

For example, a 6-bp cutter such as *Eco*RI creates about 700,000 fragments when used to digest the human genome!

One of the earliest bioinformatics applications to be developed for genomics was the use of algorithm-based software programs for creating a DNA-sequence **alignment,** in which similar sequences of bases are lined up for comparison. Alignment identifies overlapping sequences, allowing scientists to reconstruct their order in a chromosome.

Because these overlapping fragments are adjoining segments that collectively form one continuous DNA molecule within a chromosome, they are called **contiguous fragments,** or **contigs.** Figure 21.2 shows an example of contig alignment and assembly for a portion of human chromosome 2. For simplicity, this figure shows relatively short sequences for each contig, which in actuality would be much longer. The figure is also simplified in that, in actual alignments, assembled sequences do not always overlap only at their ends.

The WGS shotgun method was developed by J. Craig Venter and colleagues at The Institute for Genome Research (TIGR). In 1995, TIGR scientists used this approach to sequence the 1.83-million-bp genome of the bacterium *Haemophilus influenzae*. This was the first completed genome sequence from a free-living (i.e., nonviral) organism, and it demonstrated "proof of concept" that shotgun sequencing could be used to sequence an entire genome. Even after the genome for *H. influenzae* was sequenced, many scientists were skeptical that a shotgun approach would work on the larger genomes of eukaryotes. Now WGS approaches are the predominant method for sequencing genomes.

High-Throughput Sequencing and Its Impact on Genomics

Cutting a genome into contigs is not particularly difficult; however, a primary hurdle that had to be overcome to advance WGS was the question of how to sequence millions or billions of base pairs in a timely and cost-effective way. This was a major challenge for scientists working on the Human Genome Project (Section 21.5). The Sanger sequencing method (discussed in Chapter 20) was the predominant

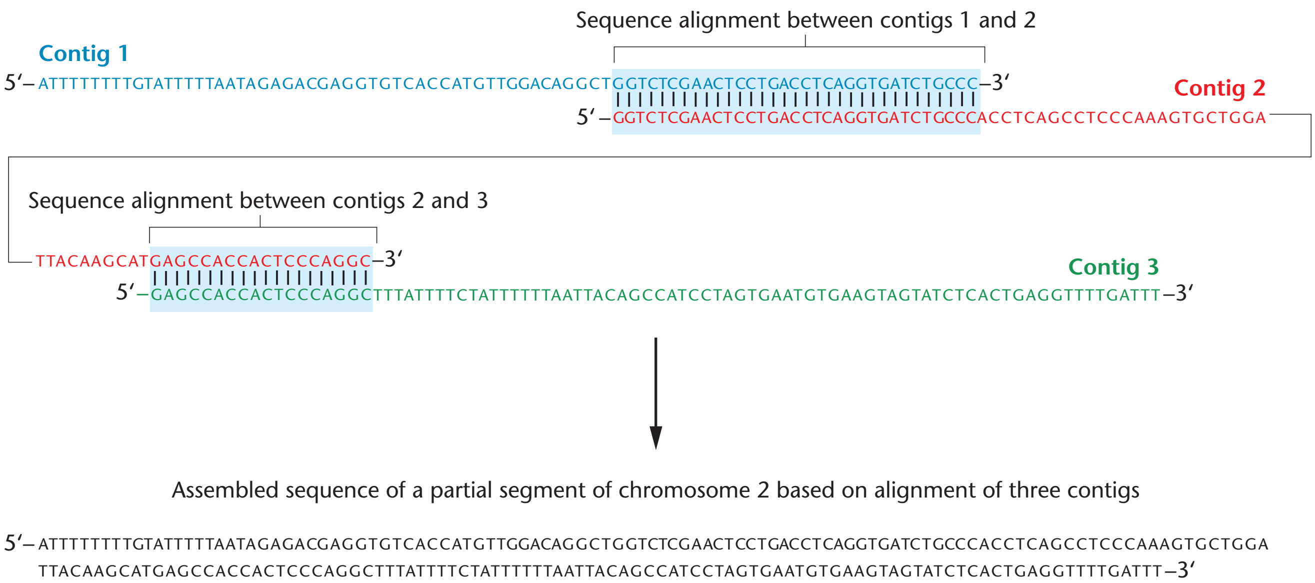

Sequence alignment between contigs 1 and 2
Contig 1
5′–ATTTTTTTTGTATTTTTAATAGAGACGAGGTGTCACCATGTTGGACAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCTGCCC–3′
5′–GGTCTCGAACTCCTGACCTCAGGTGATCTGCCCACCTCAGCCTCCCAAAGTGCTGGA
Contig 2
Sequence alignment between contigs 2 and 3
TTACAAGCATGAGCCACCACTCCCAGGC–3′
5′–GAGCCACCACTCCCAGGCTTTATTTTCTATTTTTTAATTACAGCCATCCTAGTGAATGTGAAGTAGTATCTCACTGAGGTTTTGATTT–3′
Contig 3
Assembled sequence of a partial segment of chromosome 2 based on alignment of three contigs
5′–ATTTTTTTTGTATTTTTAATAGAGACGAGGTGTCACCATGTTGGACAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCTGCCCACCTCAGCCTCCCAAAGTGCTGGA
TTACAAGCATGAGCCACCACTCCCAGGCTTTATTTTCTATTTTTTAATTACAGCCATCCTAGTGAATGTGAAGTAGTATCTCACTGAGGTTTTGATTT–3′

FIGURE 21.2 DNA-sequence alignment of contigs on human chromosome 2. Single-stranded DNA for three different contigs from human chromosome 2 is shown in blue, red, or green. The actual sequence from chromosome 2 is shown, but in reality, contig alignment involves fragments that are several thousand bases in length. Alignment of the three contigs allows a portion of chromosome 2 to be assembled. Alignment of all contigs for a particular chromosome would result in assembly of a completely sequenced chromosome.

sequencing technique for a long time. However, a major limitation of this technique was that even the best sequencing gels would typically yield only several hundred base pairs in each run and relatively few runs could be completed in a day. As a result, the overall production of sequence data was quite slow compared with modern techniques. Obviously, it would be very time consuming to manually sequence an entire genome by the Sanger method. The major technological breakthrough that made genomics possible was the development of computer-automated sequencers.

Many of the early computer-automated sequencers, designed for so-called **high-throughput sequencing,** could process millions of base pairs in a day. These sequencers contained multiple capillary gels that are each several feet long. Some ran as many as 96 capillary gels at a time, each producing around 900 bases of sequence. Because these sequencers were computer automated, they could work around the clock, generating over 2 million bases of sequence in a 24-hour period.

In the past 15 years, high-throughput sequencing has increased the productivity of DNA-sequencing technology over 500-fold. The total number of bases that could be sequenced in a single reaction was doubling about every 24 months. At the same time, this increase in efficiency brought about a dramatic decrease in cost, from about $1.00 to less than $0.001 per base pair. As we will discuss in Section 21.5, without question the development of high-throughput sequencing was essential for the Human Genome Project. And as you know from earlier in the text (see Chapter 20), next- and third-generation sequencers now enable genome scientists to produce a sequence more than 50,000 times faster than sequencers in 2000 with greater output, improved accuracy, and reduced cost.

The Clone-by-Clone Approach

Prior to the widespread use of WGS, genomes were being assembled using a **clone-by-clone approach,** also called **map-based cloning.** Initial progress on the Human Genome Project was based on this methodology, in which individual DNA fragments from restriction digests are aligned to create the restriction maps of a chromosome. These restriction fragments were then ligated into vectors such as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) to create libraries of contigs. Recall from earlier in the text (see Chapter 20) that BACs and YACs are good cloning vectors for replicating large fragments of DNA.

Prior to the development of high-throughput sequencing approaches, DNA fragments in BACs and YACs would often be further digested into smaller, more easily manipulated pieces that were then subcloned into smaller vectors such as plasmids so that they could be sequenced in their entirety. After each sequenced fragment was analyzed for alignment overlaps, a chromosome could be assembled. The bioinformatics approaches we will discuss in the next section would then be used to identify possible protein-coding genes and assign them a location on the chromosome.

Compared to WGS, the clone-by-clone approach was cumbersome and time consuming because of the time required to clone DNA fragments into different vectors, transform bacteria or yeast, select individual clones from the library for sequencing, and then carry out sequence analysis and assembly on relatively short sequences. As WGS has become a routine method for assembling genomes, map-based cloning approaches are rarely used, and only then to occasionally resolve the problems encountered during WGS.

Draft Sequences and Reference Genomes

It is common for a draft sequence of a genome to be announced before the final sequence is eventually released. But what is a final sequence, and how do scientists decide when a sequence is complete? Draft sequences often contain gaps in areas that, for any number of reasons, may have been difficult to analyze. The decision to designate a sequence as a final or **reference genome** is dictated by the degree of error that researchers are willing to assume still exists. Thus, even a reference sequence is never considered 100 percent accurate.

If a genome were only sequenced once (one read), there would be errors in the sequence. Therefore, genomes are sequenced more than once to enhance the level of accuracy. The assembly of a final or reference sequence from multiple sequencing runs is known as *compiling*. *Coverage*, or depth of sequencing, refers to the number of times a specific nucleotide appears in the same position within a sequence after multiple reads have been compiled. Next-generation sequencing approaches are considered “deep sequencing” methods because they allow for multiple reads of a sequence.

Now, a typical sequencing experiment produces billions of reads that are obtained from sequencing many different fragments of each chromosome from a biologic sample (or organism). When these reads are aligned, if a specific nucleotide appears in a specific position within the genome after multiple reads, the probability that this nucleotide represents the correct nucleotide in the sequence is higher than if different nucleotides appeared after multiple reads.

As an example, researchers using WGS on the genome of the bacterium *Pseudomonas aeruginosa* sequenced the 6.3 million nucleotides seven times to ensure that the final sequence would be accurate. Yet even with this level of redundancy, the compiling software recognized 1604 regions that required further clarification. These regions were then reanalyzed and re-sequenced. Finally, relevant parts of the shotgun sequence were compared with the sequences of two widely separated genomic regions obtained by conventional cloning. The 81,843 nucleotides cloned and sequenced by the clone-by-clone method were in perfect agreement with the sequence obtained by WGS.

Once compiled, a reference genome, the most accurate sequence available, is then analyzed to identify gene sequences, regulatory elements, and other features that reveal important information. In the next section we discuss the central role of bioinformatics in this process.

21.3 DNA Sequence Analysis Relies on Bioinformatics Applications and Genome Databases

Genomics necessitated the rapid development of **bioinformatics,** the use of computer hardware and software and mathematics applications to organize, share, and analyze data related to gene structure, gene sequence and expression, and protein structure and function. However, even before WGS projects had been initiated, a large amount of sequence information from a range of different organisms was accumulating as a result of gene cloning by recombinant DNA techniques.

Scientists around the world needed databases that could be used to store, share, and obtain the maximum amount of information from protein and DNA sequences. Thus, bioinformatics software was already being used to compare and analyze DNA sequences and to create private and public databases. But once genomics emerged as a new approach for analyzing DNA, however, bioinformatics became even more important than before. Today, it is a dynamic area of biological research, providing new career opportunities for anyone interested in merging an understanding of biological data with information technology, mathematics, and statistical analysis.

Among the most common applications of bioinformatics are to

- Compare DNA sequences, as in contig alignment, or compare sequences from different species of individuals
- Identify genes in a genomic DNA sequence
- Find gene-regulatory regions, such as promoters and enhancers
- Identify structural sequences, such as telomeric sequences, in chromosomes
- Predict the amino acid sequence of a putative polypeptide encoded by a cloned gene sequence
- Analyze protein structure, and predict protein functions on the basis of identified domains and motifs
- Deduce evolutionary relationships between genes and organisms on the basis of sequence information

High-throughput DNA-sequencing techniques were developed nearly simultaneously with the expansion of the Internet. As genome data accumulated, many DNA-sequence databases became freely available online. Databases are essential for archiving and sharing data with other researchers and with the public. One of the largest genomic databases, called **GenBank,** is maintained by the

National Center for Biotechnology Information (NCBI) in Washington, D.C., and is the largest publicly available database of DNA sequences. GenBank shares and acquires data from databases in Japan and Europe; it contains more than 220 billion bases of sequence data from over 100,000 species; and it doubles in size roughly every 18 months! The Human Genome Nomenclature Committee, supported by the NIH, establishes rules for assigning names and symbols to newly cloned human genes. As sequences are identified and genes are named, each sequence deposited into GenBank is provided with an **accession number** that scientists can use to access and retrieve that sequence for analysis.

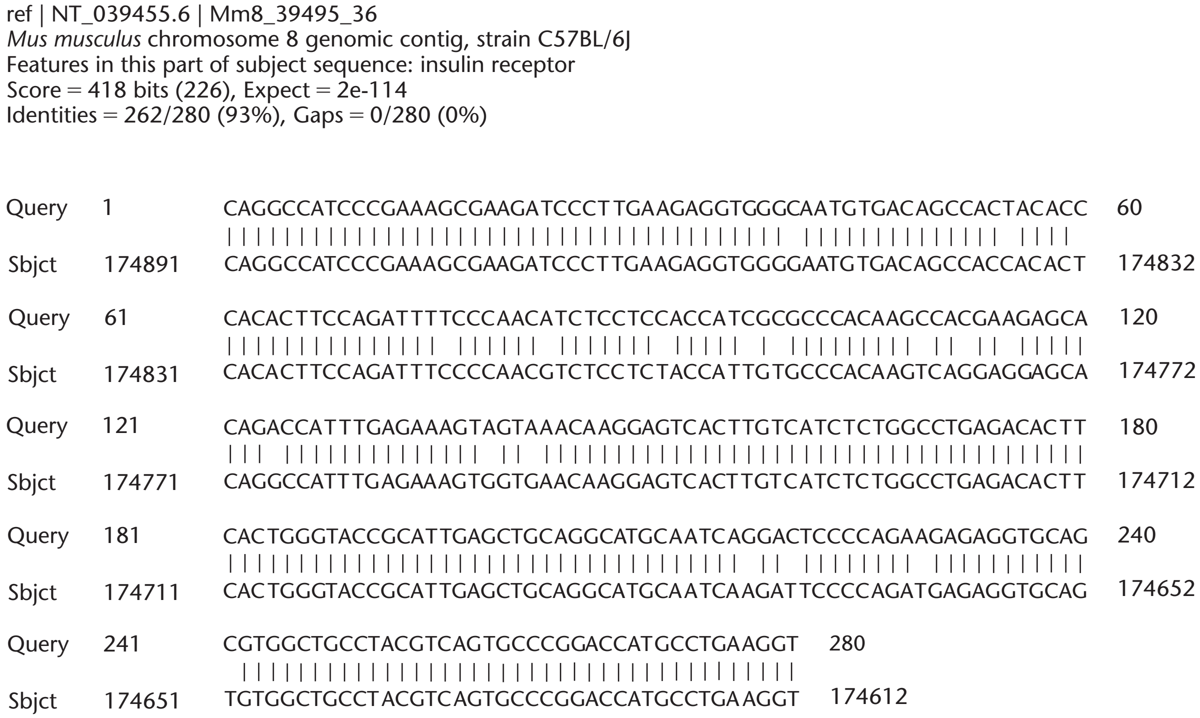
The NCBI is an invaluable source of public access databases and bioinformatics tools for analyzing genome data. You have already been introduced to NCBI and GenBank earlier in the text through several Exploring Genomics exercises. In the Exploring Genomics feature for this chapter, you will use NCBI and GenBank to compare and align contigs to assemble a chromosome segment.

Annotation to Identify Gene Sequences

One of the fundamental challenges of genomics is that, although genome projects generate tremendous amounts of DNA-sequence information, these data are of little use until they have been analyzed and interpreted. Genome projects accumulate nucleotide sequences, and then scientists have to make sense of those sequences. Thus, after a genome has been sequenced and compiled, scientists are faced with the task of identifying gene-regulatory sequences and other sequences of interest in the genome so that gene maps can be developed. This process, called **annotation**, relies heavily on bioinformatics, and a wealth of different software tools are available to carry it out.

One initial approach to annotating a sequence is to compare the newly sequenced genomic DNA to the known sequences already stored in various databases. The NCBI provides access to **BLAST (Basic Local Alignment Search Tool),** a very popular software application for searching through banks of DNA and protein sequence data. Using BLAST, we can compare a segment of genomic DNA to sequences throughout major databases such as GenBank to identify portions that align with or are the same as existing sequences. For WGS projects, simple BLAST alignments are insufficient and more complex algorithms are required to align the billions of reads of DNA sequence generated.

Figure 21.3 shows a representative example of a sequence alignment based on a BLAST search. Here a 280-bp contig from rat chromosome 12 was used to search



```
ref | NT_039455.6 | Mm8_39495_36
Mus musculus chromosome 8 genomic contig, strain C57BL/6J
Features in this part of subject sequence: insulin receptor
Score = 418 bits (226), Expect = 2e-114
Identities = 262/280 (93%), Gaps = 0/280 (0%)


Query   1        CAGGCCATCCCGAAAGCGAAGATCCCTTGAAGAGGTGGGCAATGTGACAGCCACTACACC   60
                 ||||||||||||||||||||||||||||||||||||||| ||||||||||||| ||||
Sbjct   174891   CAGGCCATCCCGAAAGCGAAGATCCCTTGAAGAGGTGGGGAATGTGACAGCCACCACACT   174832

Query   61       CACACTTCCAGATTTTCCCAACATCTCCTCCACCATCGCGCCCACAAGCCACGAAGAGCA   120
                 ||||||||||||||| |||||| ||||||| ||||| | ||||||||| || || |||||
Sbjct   174831   CACACTTCCAGATTTCCCCAACGTCTCCTCTACCATTGTGCCCACAAGTCAGGAGGAGCA   174772

Query   121      CAGACCATTTGAGAAAGTAGTAAACAAGGAGTCACTTGTCATCTCTGGCCTGAGACACTT   180
                 ||| |||||||||||||| || ||||||||||||||||||||||||||||||||||||||
Sbjct   174771   CAGGCCATTTGAGAAAGTGGTGAACAAGGAGTCACTTGTCATCTCTGGCCTGAGACACTT   174712

Query   181      CACTGGGTACCGCATTGAGCTGCAGGCATGCAATCAGGACTCCCCAGAAGAGAGGTGCAG   240
                 |||||||||||||||||||||||||||||||||||| || |||||||| |||||||||||
Sbjct   174711   CACTGGGTACCGCATTGAGCTGCAGGCATGCAATCAAGATTCCCCAGATGAGAGGTGCAG   174652

Query   241      CGTGGCTGCCTACGTCAGTGCCCGGACCATGCCTGAAGGT   280
                  |||||||||||||||||||||||||||||||||||||||
Sbjct   174651   TGTGGCTGCCTACGTCAGTGCCCGGACCATGCCTGAAGGT   174612
```

FIGURE 21.3 BLAST results showing a 280-base sequence of a chromosome 12 contig from rats (*Rattus norvegicus*, the "query") aligned with a portion of chromosome 8 from mice (*Mus musculus*, the "subject") that contains a partial sequence for the insulin receptor gene. Vertical lines indicate exact matches. The rat contig sequence was used as a query sequence to search a mouse database in GenBank. Notice that the two sequences show 93 percent identity, strong evidence that this rat contig sequence contains a gene for the insulin receptor.

a mouse database to determine whether a sequence in the rat contig matched a known gene in mice. Notice that the rat contig (the query sequence in the BLAST search) aligned with base pairs 174,612 to 174,891 of mouse chromosome 8. The accession number for the mouse chromosome sequence, NT_039455.6, is indicated at the top of the figure. BLAST searches calculate an **identity value**—determined by the sum of identical matches between aligned sequences divided by the total number of bases aligned. Gaps, indicating missing bases in the two sequences, are usually ignored in calculating similarity scores. The aligned rat and mouse sequences were 93 percent identical and showed no gaps in the alignment.

Notice that the BLAST report also provides an "Expect" value, or **E-value,** based on the number of matching sequences in the database that would be expected by chance. E-values take into account the length of the query sequence. Shorter sequences have a much greater likelihood of being present in the database by chance than longer sequences. The lower the E-value (the closer it is to 0), the higher the significance of the match. Significant alignments, indicating that DNA sequences are significantly similar, have E-values much less than 1.0.

Because this mouse sequence on chromosome 8 is known to contain an insulin receptor gene (encoding a protein that binds the hormone insulin), it is highly likely that the rat contig sequence also contains an insulin receptor gene. We will return to the topic of similarity in Sections 21.4 and 21.7, where we consider how similarity between gene sequences can be used to infer function and to identify evolutionarily related genes through comparative genomics.

Hallmark Characteristics of a Gene Sequence Can Be Recognized during Annotation

A major limitation of a BLAST search for annotation is that it only works if similar gene sequences are already in a database. Fortunately, BLAST is not the only way to identify genes. Whether the genome under study is from a eukaryote or a bacterium, several hallmark characteristics of genes can be searched for using bioinformatics software (**Figure 21.4**). We discussed many of these characteristics of a "typical" gene earlier in the text (see Chapters 13 and 17).

For instance, gene-regulatory sequences found upstream of genes are marked by identifiable sequences such as promoters, enhancers, and silencers. Recall from earlier in the text (see Chapter 17) that TATA box, GC box, and CAAT box sequences are often present in the promoter region of eukaryotic genes.

Recall also that splice sites between **exons** and **introns** contain a predictable sequence (most introns begin with CT and end with AG) and such splice-site sequences are important for determining intron and exon boundaries. Annotation can sometimes be a little bit easier for bacterial genes than for eukaryotic genes because there are no introns in bacterial genes.

Downstream elements, such as termination sequences and well-defined sequences at the end of a gene, where a polyadenylation sequence signals the addition of a poly-A tail to the 3′ end of an mRNA transcript, are also important for annotation.

Gene-prediction programs are used to annotate sequences. These programs incorporate search elements for many of the characteristics noted in figure 21.4 and have become invaluable applications of bioinformatics. Yet even with bioinformatics, identifying a gene in a particular sequence of DNA is not always straightforward, particularly when one is studying genes that do not code for proteins. In fact, a reasonable question whenever one sequences a genome is, "Where are the genes?" In other words, how does one know what sequences of a genome are genes and which sequences are not genes or parts of a gene?

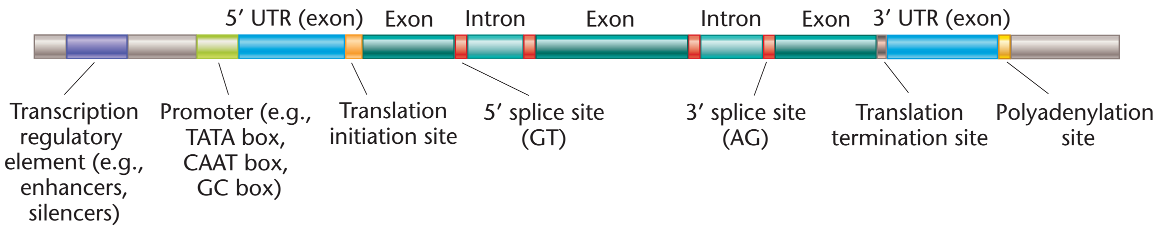
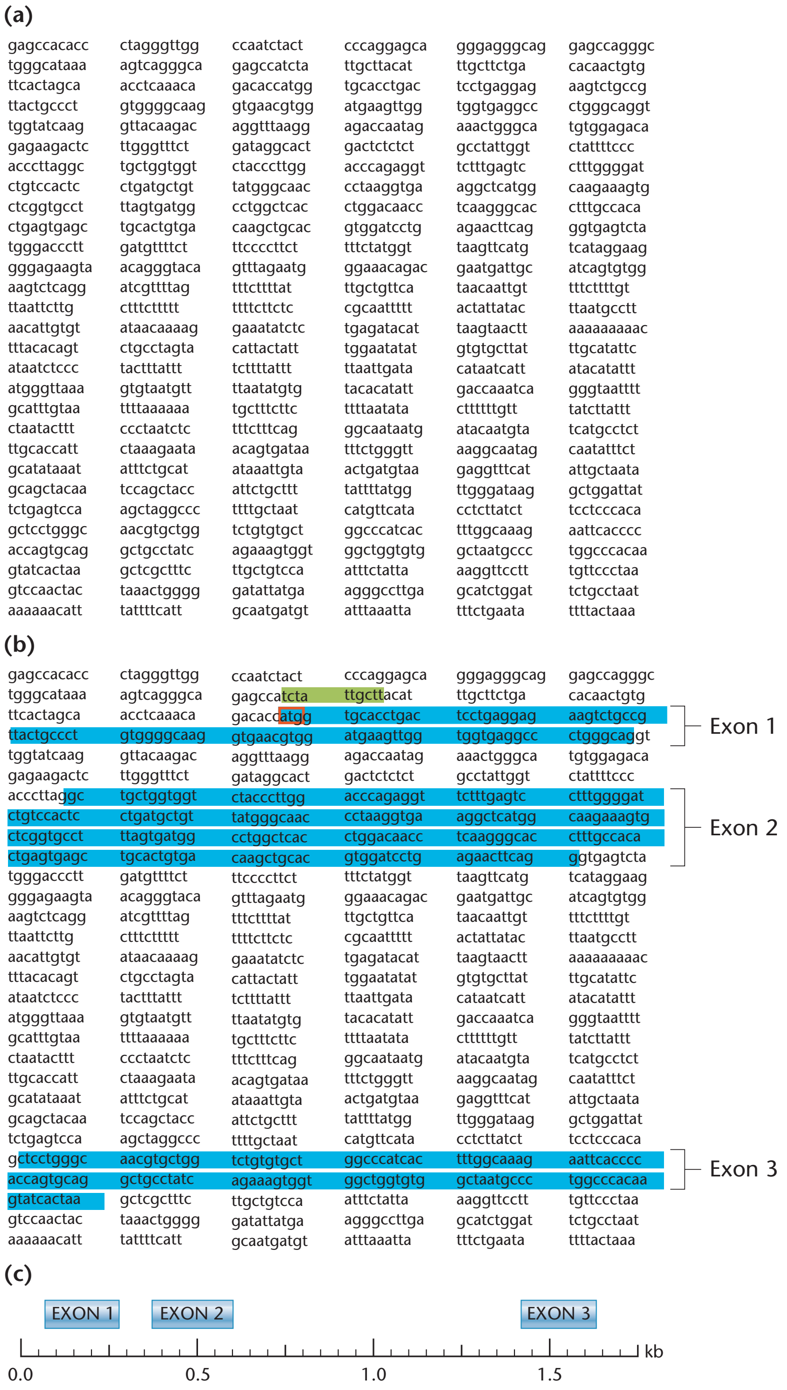

5′ UTR (exon)
Exon
Intron
Exon
Intron
Exon
3′ UTR (exon)
Transcription regulatory element (e.g., enhancers, silencers)
Promoter (e.g., TATA box, CAAT box, GC box)
Translation initiation site
5′ splice site (GT)
3′ splice site (AG)
Translation termination site
Polyadenylation site

FIGURE 21.4 Characteristics of a protein-coding gene that can be used during annotation to identify a gene in an unknown sequence of genomic DNA. Most eukaryotic genes are organized into coding segments (exons) and noncoding segments (introns). When annotating a genome sequence to determine whether it contains a gene, it is necessary to distinguish between introns and exons, gene-regulatory sequences, such as promoters and enhancers, untranslated regions (UTRs), and gene termination sequences.



(a)

gagccacacc	ctagggttgg	ccaatctact	cccaggagca	gggagggcag	gagccagggc
tgggcataaa	agtcagggca	gagccatcta	ttgcttacat	ttgcttctga	cacaactgtg
ttcactagca	acctcaaaca	gacaccatgg	tgcacctgac	tcctgaggag	aagtctgccg
ttactgccct	gtggggcaag	gtgaacgtgg	atgaagttgg	tggtgaggcc	ctgggcaggt
tggtatcaag	gttacaagac	aggtttaagg	agaccaatag	aaactgggca	tgtggagaca
gagaagactc	ttgggtttct	gataggcact	gactctctct	gcctattggt	ctattttccc
acccttaggc	tgctggtggt	ctacccttgg	acccagaggt	tctttgagtc	ctttggggat
ctgtccactc	ctgatgctgt	tatgggcaac	cctaaggtga	aggctcatgg	caagaaagtg
ctcggtgcct	ttagtgatgg	cctggctcac	ctggacaacc	tcaagggcac	ctttgccaca
ctgagtgagc	tgcactgtga	caagctgcac	gtggatcctg	agaacttcag	ggtgagtcta
tgggaccctt	gatgttttct	ttccccttct	tttctatggt	taagttcatg	tcataggaag
gggagaagta	acagggtaca	gtttagaatg	ggaaacagac	gaatgattgc	atcagtgtgg
aagtctcagg	atcgttttag	tttcttttat	ttgctgttca	taacaattgt	tttcttttgt
ttaattcttg	ctttcttttt	ttttcttctc	cgcaattttt	actattatac	ttaatgcctt
aacattgtgt	ataacaaaag	gaaatatctc	tgagatacat	taagtaactt	aaaaaaaaac
tttacacagt	ctgcctagta	cattactatt	tggaatatat	gtgtgcttat	ttgcatattc
ataatctccc	tactttattt	tcttttattt	ttaattgata	cataatcatt	atacatattt
atgggttaaa	gtgtaatgtt	ttaatatgtg	tacacatatt	gaccaaatca	gggtaatttt
gcatttgtaa	ttttaaaaaa	tgctttcttc	ttttaatata	cttttttgtt	tatcttattt
ctaatacttt	ccctaatctc	tttctttcag	ggcaataatg	atacaatgta	tcatgcctct
ttgcaccatt	ctaaagaata	acagtgataa	tttctgggtt	aaggcaatag	caatatttct
gcatataaat	atttctgcat	ataaattgta	actgatgtaa	gaggtttcat	attgctaata
gcagctacaa	tccagctacc	attctgcttt	tattttatgg	ttgggataag	gctggattat
tctgagtcca	agctaggccc	ttttgctaat	catgttcata	cctcttatct	tcctcccaca
gctcctgggc	aacgtgctgg	tctgtgtgct	ggcccatcac	tttggcaaag	aattcacccc
accagtgcag	gctgcctatc	agaaagtggt	ggctggtgtg	gctaatgccc	tggcccacaa
gtatcactaa	gctcgctttc	ttgctgtcca	atttctatta	aaggttcctt	tgttccctaa
gtccaactac	taaactgggg	gatattatga	agggccttga	gcatctggat	tctgcctaat
aaaaaacatt	tattttcatt	gcaatgatgt	atttaaatta	tttctgaata	ttttactaaa

(b)

gagccacacc	ctagggttgg	ccaatctact	cccaggagca	gggagggcag	gagccagggc	
tgggcataaa	agtcagggca	gagccatcta	ttgcttacat	ttgcttctga	cacaactgtg	
ttcactagca	acctcaaaca	gacaccatgg	tgcacctgac	tcctgaggag	aagtctgccg	Exon 1
ttactgccct	gtggggcaag	gtgaacgtgg	atgaagttgg	tggtgaggcc	ctgggcaggt	
tggtatcaag	gttacaagac	aggtttaagg	agaccaatag	aaactgggca	tgtggagaca	
gagaagactc	ttgggtttct	gataggcact	gactctctct	gcctattggt	ctattttccc	
acccttaggc	tgctggtggt	ctacccttgg	acccagaggt	tctttgagtc	ctttggggat	Exon 2
ctgtccactc	ctgatgctgt	tatgggcaac	cctaaggtga	aggctcatgg	caagaaagtg	
ctcggtgcct	ttagtgatgg	cctggctcac	ctggacaacc	tcaagggcac	ctttgccaca	
ctgagtgagc	tgcactgtga	caagctgcac	gtggatcctg	agaacttcag	ggtgagtcta	
tgggaccctt	gatgttttct	ttccccttct	tttctatggt	taagttcatg	tcataggaag	
gggagaagta	acagggtaca	gtttagaatg	ggaaacagac	gaatgattgc	atcagtgtgg	
aagtctcagg	atcgttttag	tttcttttat	ttgctgttca	taacaattgt	tttcttttgt	
ttaattcttg	ctttcttttt	ttttcttctc	cgcaattttt	actattatac	ttaatgcctt	
aacattgtgt	ataacaaaag	gaaatatctc	tgagatacat	taagtaactt	aaaaaaaaac	
tttacacagt	ctgcctagta	cattactatt	tggaatatat	gtgtgcttat	ttgcatattc	
ataatctccc	tactttattt	tcttttattt	ttaattgata	cataatcatt	atacatattt	
atgggttaaa	gtgtaatgtt	ttaatatgtg	tacacatatt	gaccaaatca	gggtaatttt	
gcatttgtaa	ttttaaaaaa	tgctttcttc	ttttaatata	cttttttgtt	tatcttattt	
ctaatacttt	ccctaatctc	tttctttcag	ggcaataatg	atacaatgta	tcatgcctct	
ttgcaccatt	ctaaagaata	acagtgataa	tttctgggtt	aaggcaatag	caatatttct	
gcatataaat	atttctgcat	ataaattgta	actgatgtaa	gaggtttcat	attgctaata	
gcagctacaa	tccagctacc	attctgcttt	tattttatgg	ttgggataag	gctggattat	
tctgagtcca	agctaggccc	ttttgctaat	catgttcata	cctcttatct	tcctcccaca	
gctcctgggc	aacgtgctgg	tctgtgtgct	ggcccatcac	tttggcaaag	aattcacccc	Exon 3
accagtgcag	gctgcctatc	agaaagtggt	ggctggtgtg	gctaatgccc	tggcccacaa	
gtatcactaa	gctcgctttc	ttgctgtcca	atttctatta	aaggttcctt	tgttccctaa	
gtccaactac	taaactgggg	gatattatga	agggccttga	gcatctggat	tctgcctaat	
aaaaaacatt	tattttcatt	gcaatgatgt	atttaaatta	tttctgaata	ttttactaaa	

FIGURE 21.5 Annotation of a cDNA sequence containing part of the human β-globin gene. By convention, the sequence is presented in groups of ten nucleotides, although in reality the sequence is continuous. (a) The location of genes, if any, in this sequence is not readily apparent from a cursory glance. (b) The analyzed sequence, showing the location of a promoter sequence (green). Open reading frames for three exons of the gene are shown in blue. (c) Diagrammatic representation of three exons for the human β-globin gene encoded by the sequence shown in (a) and (b).

Consider the sequence presented in Figure 21.5(a), which shows a portion of the human genome. From a casual inspection, it is not clear whether this sequence contains any genes and, if so, how many. Analysis of the sequence, however, may reveal characteristics (like those reviewed above and summarized in Figure 21.4) that provide clues to the presence of a protein-coding gene. In addition, protein-coding genes contain one or more **open reading frames (ORFs),** nucleotide sequences that, after transcription and mRNA splicing, are translated into the amino acid sequence of a protein.

ORFs typically begin with an initiation sequence—usually ATG, which transcribes into the AUG start codon of an mRNA molecule—and end with a termination sequence—TAA, TAG, or TGA—which corresponds to the stop codons of UAA, UAG, and UGA in mRNA. Genetic information is encoded in groups of three nucleotides (triplets), but it is not always clear whether to begin the analysis of a sequence at the first nucleotide, the second, or the third. Typically, the sequence adjacent to a promoter is examined for a start (initiation) triplet; however, ORFs can be used to identify a gene even when a promoter sequence is not apparent. Software programs can then analyze the ORFs three nucleotides at a time. The discovery of an ORF starting with an ATG followed at some distance by a termination sequence is usually a good indication that the coding region of a gene has been identified.

The way genes are organized in eukaryotic genomes (including the human genome) makes direct searching for ORFs more difficult in eukaryotic genomes than in bacterial genomes. First, many eukaryotic genes have introns. As a result, many, if not most, eukaryotic genes are not organized as continuous ORFs; instead, the gene sequences consist of ORFs (exons) interspersed with introns. Second, genes in humans and other eukaryotes are often widely spaced, increasing the chances of finding false ORFs in the regions between gene clusters.

Now look back at the sequence in Figure 21.5(a). The sequence selected for this example is NM_000518.4 *Homo sapiens* hemoglobin subunit beta (HBB), mRNA positive strand (non-template strand). In fact, annotation reveals several identifiable indicators that the sequence contains a protein-coding gene: It includes a promoter sequence, an initiation triplet, and three exons [Figure 21.5(b)]. The two unshaded regions between the exons represent introns that would be spliced out following transcription when the mRNA is processed. Using this sequence as the query in a search of genomic databases would reveal that it is part of a single gene, the human β-globin gene.

Software designed for ORF analysis of eukaryotic genomes is highly valuable. Often such programs are used to make computational predictions of all ORFs (the ORFeome) in a sequenced genome as a way to estimate its number of potential protein-coding genes. Thus annotation can be used to predict the number of proteins encoded by a genome. In addition to the features already mentioned, such software can also be used to "translate" ORFs into possible polypeptide sequences as a way to predict the polypeptide encoded by a gene. For example, **Figure 21.6** shows a partial sequence for the first exon of the human tubulin alpha 3c gene (*TUBA3C*). Prediction programs scan potential ORFs in the 5′ to 3′ direction on both strands of a section of genomic DNA to predict possible reading frames in each direction. Figure 21.6 shows the results for the six possible reading frames in the sequence of interest. Amino acids are shown using the single-letter code for each residue. Notice the very different results obtained for each of the six frames. For instance, the 5′ to 3′ ORF 1 contains several stop codons interspersed among amino acids but no methionine residues that are evidence of a start codon. Other ORFs would contain too many methionines to produce a functional polypeptide. For this exon of *TUBA3C*, the 5′ to 3′ ORF 2 is the correct reading frame, but this would not be obvious just based on a visual inspection of the sequence or even just through bioinformatics. That this reading frame is correct has been confirmed by experimental methods—as is often necessary even when bioinformatics predicts possible ORFs.

Prediction programs can also search for **codon bias,** the more frequent use of one or two codons to encode an amino acid that can be specified by a number of different codons. For example, alanine can be encoded by GCA, GCT, GCC, and GCG. If the codons were used randomly, each would be used about 25 percent of the time. Yet in the human genome, GCC is used 41 percent of the time, and GCG only 11 percent of the time. Codon bias is present in exons but should not be present in introns or intergenic spacers.

NOW SOLVE THIS

21.1 In a sequence encompassing 99.4 percent of the euchromatic regions of human chromosome 1, Gregory et al. [(2006) *Nature* 441:315–321] identified 3141 genes.

(a) How does one identify a gene within a raw sequence of bases in DNA?

(b) What features of a genome are used to verify likely gene assignments?

(c) Given that chromosome 1 contains approximately 8 percent of the human genome, and assuming that there are approximately 20,000 genes, would you consider chromosome 1 to be "gene rich"?

■ **HINT:** *This problem involves a basic understanding of bioinformatics and gene annotation approaches to determine how potential gene sequences can be identified in a stretch of sequenced DNA.*

(a) *Homo sapiens TUBA3C* (bp 1–300)

```
  1 ggttgaggtcaagtagtagcgttgggctgcggcagcggaggagctcaacatgcgtgagtg
 61 tatctctatccacgtggggcaggcaggagtccagatcggcaatgcctgctgggaactgta
121 ctgcctggaacatggaattcagcccgatggtcagatgccaagtgataaaaccattggtgg
181 tggggacgactccttcaacacgttcttcagtgagactggagctggcaagcacgtgcccag
241 agcagtgtttgtggacctggagcccactgtggtcgatgaagtgcgcacaggaacctatag (300)
```

(b) Predicted polypeptides

5' to 3' Frame 1

G **Stop** G Q V V A L G C G S G G A Q H A **Stop** V Y L Y P R G A G R S P D R Q C L L G T V L P G T W N S A R W S D A K **Stop** **Stop** N H W W W G R L L Q H V L Q **Stop** D W S W Q A R A Q S S V C G P G A H C G R **Stop** S A H R N L **Stop**

5' to 3' Frame 2

V E V K **Stop** **Stop** R W A A A A E E L N **Met** R E C I S I H V G Q A G V Q I G N A C W E L Y C L E H G I Q P D G Q **Met** P S D K T I G G G D D S F N T F F S E T G A G K H V P R A V F V D L E P T V V D E V R T G T Y

5' to 3' Frame 3

L R S S S S V G L R Q R R S S T C V S V S L S T W G R Q E S R S A **Met** P A G N C T A W N **Met** E F S P **Met** V R C Q V I K P L V V G T T P S T R S S V R L E L A S T C P E Q C L W T W S P L W S **Met** K C A Q E P I

3' to 5' Frame 1

L **Stop** V P V R T S S T T V G S R S T N T A L G T C L P A P V S L K N V L K E S S P P P **Met** V L S L G I **Stop** P S G **Stop** I P C S R Q Y S S Q Q A L P I W T P A C P T W I E I H S R **Met** L S S S A A A A Q R Y Y L T S T

3' to 5' Frame 2

Y R F L C A L H R P Q W A P G P Q T L L W A R A C Q L Q S H **Stop** R T C **Stop** R S R P H H Q W F Y H L A S D H R A E F H V P G S T V P S R H C R S G L L P A P R G **Stop** R Y T H A C **Stop** A P P L P Q P N A T T **Stop** P Q

3' to 5' Frame 3

I G S C A H F I D H S G L Q V H K H C S G H V L A S S S L T E E R V E G V V P T T N G F I T W H L T I G L N S **Met** F Q A V Q F P A G I A D L D S C L P H V D R D T L T H V E L L R C R S P T L L L D L N

FIGURE 21.6 Predicted polypeptide sequences translated from potential ORFs in the human *TUBA3C* gene derived from the non-coding (positive) strand. (a) Nucleotides 1–300 of the first exon in the human *TUBA3C* gene. (b) A translation program predicts six possible polypeptide sequences from this exon. Which predicted sequence is correct?

Note: Methionine is highlighted using the three-letter amino acid code (Met).

21.4 Functional Genomics Establishes Gene Function and Identifies Regulatory Elements in a Genome

Functional genomics interprets DNA sequences and establishes gene functions based on the projected RNAs or possible proteins they encode and, as well, identifies other components of the genome, such as gene-regulatory elements. Functional genomics can involve experimental approaches to confirm or refute computational predictions (such as the number of protein-coding genes).

Predicting Gene and Protein Functions by Sequence Analysis

One approach to assigning functions to genes is to use *sequence similarity searches*, as described in the previous section. Programs such as BLAST are used to search through databases to find alignments between the newly sequenced genome and genes that have already been identified, either in the same or in different species. You were introduced to this approach for predicting gene function in Figure 21.3. Inferring gene function from similarity searches is based on a relatively simple idea. If a genome sequence shows statistically significant similarity to the sequence of a gene whose function is known, then it is likely that the genome sequence encodes a protein with a similar or related function.

Another major benefit of similarity searches is that they are often able to identify **homologous genes,** genes that are evolutionarily related.

Homologous genes in the same species are called **paralogs**. In the globin gene family, the α- and β-globin subunits in humans are paralogs resulting from a gene-duplication event. Paralogs often have similar or identical functions.

If homologous genes in different species are thought to have descended from a gene in a common ancestor, the genes are known as **orthologs.** For instance, mouse and human α-globin genes are orthologs evolved from a common ancestor. After the human genome was sequenced, many ORFs in it were identified as protein-coding genes based on their alignment with related genes of known function in other species. As an example, **Figure 21.7** compares portions of the human leptin gene (*LEP*) with its ortholog in mice (*Lep*). These two genes are over 85 percent identical in sequence. The leptin gene was first discovered in mice. The match between the *LEP*-containing DNA sequence in humans and the mouse ortholog sequence confirms the identity and leptin-coding function of this gene in human genomic DNA.

As an interesting aside, the leptin gene (also called *Lep, for leptin*, in mice) is highly expressed in fat cells (adipocytes). This gene produces the protein hormone leptin, which targets cells in the brain to suppress appetite. Knockout mice lacking a functional *Lep* gene grow dramatically overweight [see Figure 20.18(b)]. A similar phenotype has been observed in small numbers of humans with particular mutations in *LEP.* Although it is important to note that weight control is not regulated by a single gene, the discovery of leptin has provided significant insight into lipid metabolism and weight disorders in humans.

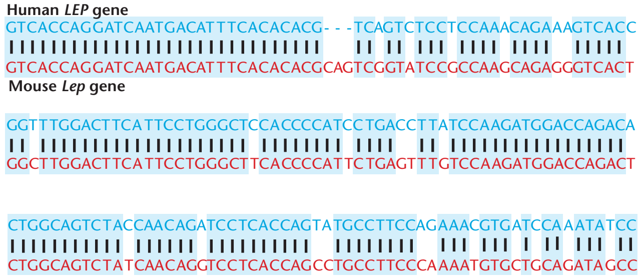

Human *LEP* gene
GTCACCAGGATCAATGACATTTCACACACG- - -TCAGTCTCCTCCAAACAGAAAGTCACC
GTCACCAGGATCAATGACATTTCACACACGCAGTCGGTATCCGCCAAGCAGAGGGTCACT
Mouse *Lep* gene
GGTTTGGACTTCATTCCTGGGCTCCACCCCATCCTGACCTTATCCAAGATGGACCAGACA
GGCTTGGACTTCATTCCTGGGCTTCACCCCATTCTGAGTTTGTCCAAGATGGACCAGACT
CTGGCAGTCTACCAACAGATCCTCACCAGTATGCCTTCCAGAAACGTGATCCAAATATCC
CTGGCAGTCTATCAACAGGTCCTCACCAGCCTGCCTTCCCAAAATGTGCTGCAGATAGCC

FIGURE 21.7 Comparison of the human *LEP* and mouse *Lep* genes. Partial sequences for these orthologs are shown with the human *LEP* gene on top (in blue) and the mouse *Lep* gene sequence below it (in red). Notice from the number of identical nucleotides, indicated by shaded boxes and vertical lines, that the nucleotide sequence for these two genes is very similar.

Predicting Function from Structural Analysis of Protein Domains and Motifs

When a gene sequence is used to predict a polypeptide sequence, the polypeptide sequence can be analyzed for specific structural domains and motifs. Identification of **protein domains,** such as ion channels, membrane-spanning regions, DNA-binding regions, secretion and export signals, and other structural aspects of a polypeptide that are encoded by a DNA sequence, can in turn be used to predict protein function. Recall from earlier in the text (see Chapter 17), for example, that the structures of many DNA-binding proteins have characteristic patterns, or **motifs,** such as the helix-turn-helix, leucine zipper, or zinc-finger motifs. These motifs can often easily be searched for using bioinformatics software, and their identification in a sequence is a common strategy for inferring the possible functions of a protein.

Investigators Are Using Genomics Techniques Such as Chromatin Immunoprecipitation to Investigate Aspects of Genome Function and Regulation

In this chapter and later in the text (see Chapter 22), we will consider a range of different genomic techniques that are valuable for functional genomics studies. These include various methods designed to map protein—DNA interactions and which are also useful for identifying genes that are regulated by DNA-binding transcription factors. Recall from earlier in the text (see Chapter 13) that transcription factors bind to specific sequences in the genome adjacent to genes to help initiation transcription by RNA polymerases. One example is a technique called **chromatin immunoprecipitation (ChIP)** (**Figure 21.8**). There are different approaches to ChIP, but the end result of this method is the recovery of DNA fragments attached to DNA-binding proteins. Investigators can then purify the immunoprecipitated DNA fragments by removing them from the antibody/protein complex.

There are several options for analyzing such precipitated DNA fragments. Figure 21.8 specifically demonstrates an approach called *ChIP sequencing*, or *ChIPSeq*. Here the fragments are directly sequenced by high-throughput approaches. This allows researchers to study an entire genome to locate binding sites for proteins such as transcription factors, histone-related proteins, and other proteins involved in chromatin structure.

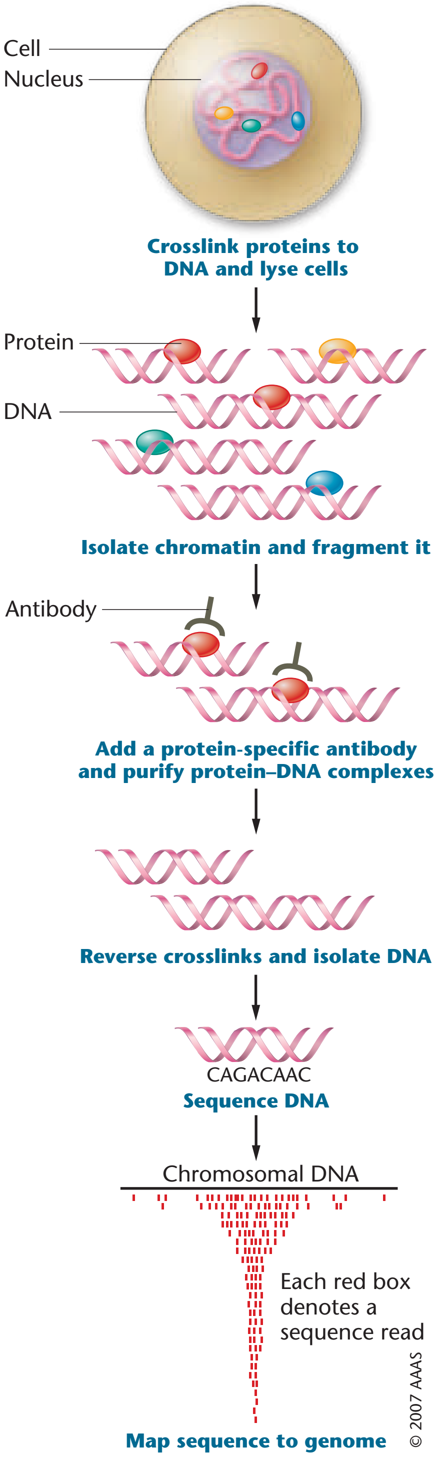

Cell
Nucleus
Crosslink proteins to DNA and lyse cells
Protein
DNA
Isolate chromatin and fragment it
Antibody
Add a protein-specific antibody and purify protein–DNA complexes
Reverse crosslinks and isolate DNA
CAGACAAC
Sequence DNA
Chromosomal DNA
Each red box denotes a sequence read
Map sequence to genome
© 2007 AAAS

FIGURE 21.8 The ChIPSeq method screens for specific transcription factor binding sites across a whole genome. In this method, formaldehyde is added to tissues or cultured cells to crosslink DNA-binding proteins currently attached to chromatin. Then the chromatin is extracted from cells and sheared into small fragments. An antibody or antibodies that recognize specific DNA-binding proteins of interest (POI), such as a transcription factor, are added to the mixture, and the antibodies attach to the POI. Then the antibody, together with its protein–DNA fragment, is pulled out of the mixture (immunoprecipitated) by centrifugation. The immunoprecipitated DNA fragments are released from crosslinked proteins and attached antibodies and are then sequenced. Sequence data reveal the DNA-binding site for the POI and these sequences can be mapped to specific locations in the genome.

21.5 The Human Genome Project Revealed Many Important Aspects of Genome Organization in Humans

Now that you have a general idea of the basic strategies used for analyzing a genome, let's look at the largest genomics project completed to date. The **Human Genome Project (HGP)** was a coordinated international effort to determine the sequence of the human genome and to identify all the genes it contains. It has produced a plethora of information, much of which is still being analyzed and interpreted. What is already clear, based on all the different kinds of genomes that have been sequenced, is that humans and all other species share a common set of genes essential for cellular function and reproduction, confirming that all living organisms arose from a common ancestor.

Origins of the Project

The publicly funded Human Genome Project began in 1990 under the direction of James Watson, the co-discoverer of the double-helix structure of DNA. Eventually the public project was led by Dr. Francis Collins, who had previously led a research team involved in identifying the *CFTR* gene as the cause of cystic fibrosis. In the United States, the Collins-led HGP was coordinated by the Department of Energy and the National Center of Human Genome Research, a division of the National Institutes of Health. It established a 15-year plan with a proposed budget of $3 billion to identify all human genes, originally thought to number between 80,000 and 100,000, to sequence and map them all, and to sequence the approximately 3 billion base pairs thought to be comprised by the 24 chromosomes (22 autosomes, plus X and Y) in humans. Other primary goals of the HGP included the following:

- Establish functional categories for all human genes.
- Analyze genetic variations between humans, including the identification of single-nucleotide polymorphisms (SNPs).
- Map and sequence the genomes of several model organisms used in experimental genetics, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus* (mouse).
- Develop new sequencing technologies, such as high-throughput computer-automated sequencers, to facilitate genome analysis.
- Disseminate genome information among both scientists and the general public.

Recognizing the impact that genetic information would have on society, the HGP also set up the **ELSI program** (standing for Ethical, Legal, and Social Implications) to consider ethical, legal, and social issues arising from the HGP and to ensure that personal genetic information would be safeguarded and not used in discriminatory ways.

As the HGP grew into an international effort, scientists in 18 countries were involved in the project. Much of the work was carried out by the International Human Genome Sequence Consortium, involving nearly 3000 scientists working at 20 centers in six countries (China, France, Germany, Great Britain, Japan, and the United States).

In 1999, a privately funded human genome project led by J. Craig Venter at *Celera Genomics* (aptly named from a word meaning "swiftness") was announced. Celera's goal was to use WGS and computer-automated high-throughput DNA sequencers to sequence the human genome more rapidly than HGP. The public project had proposed using a clone-by-clone approach to sequence the genome. (Recall that Venter and his colleagues had proven the potential of WGS in 1995 when they completed the genome for *H. influenzae*.) Celera's announcement set off an intense competition between the two teams, which both aspired to be first with the human genome sequence. This contest eventually led to the HGP finishing ahead of schedule and under budget after scientists from the public project began to use high-throughput sequencers and WGS strategies as well.

Major Features of the Human Genome

In June 2000, the leaders of the public and private genome projects met at the White House with President Bill Clinton and jointly announced the completion of a draft sequence of the human genome. In February 2001, they each published an analysis covering about 96 percent of the euchromatic region of the genome. The public project sequenced euchromatic portions of the genome 12 times and set a quality-control standard of a 0.01 percent error rate for their sequence. Although this error rate may seem very low, it still allows about 600,000 errors in the human genome sequence. Celera sequenced certain areas of the genome more than 35 times when compiling the genome.

The remaining work of completing the sequence involved filling in gaps clustered around centromeres, telomeres, and repetitive sequences (regions rich in GC base pairs can be particularly tough to sequence and interpret), correcting misaligned segments, and re-sequencing portions of the genome to ensure accuracy. In 2003 sequencing and error fixing were deemed sufficient to pass the international project's definition of completion—that the analysis contained fewer than 1 error per 10,000 nucleotides and that it covered 95 percent of the gene-containing portions of the genome. Yet even at the time of "completion" there were still some 350 gaps in the sequence that continued to be worked on.

And obviously the HGP did not sequence the genome of every person on Earth. The assembled sequence consists of haploid genomes pooled from different individuals so that they provide a reference genome representative of major, common elements of a human genome widely shared among populations of humans. Examples of major features of the human genome are summarized in **Table 21.1**. As you can see in this table, many unexpected observations have provided us with major new insights. The genome is not static! Genome variations, including the abundance of repetitive sequences scattered throughout the genome, verify that the genome is dynamic, revealing many evolutionary examples of sequences that have changed in structure and location. In many ways, the HGP has revealed just how little we know about our genome.

Two of the biggest surprises discovered by the HGP were that less than 2 percent of the genome codes for proteins and

TABLE 21.1 Major Features of the Human Genome

- The human genome contains ~3.1 billion nucleotides, but protein-coding sequences make up only about 2 percent of the genome.
- The genome sequence is ~99.9 percent similar in individuals of all nationalities. SNPs and copy number variations (CNVs) account for genome diversity from person to person.
- The genome is dynamic. At least 50 percent of the genome is derived from transposable elements, such as LINE and *Alu* sequences, and other repetitive DNA sequences.
- The human genome contains approximately 20,000 protein-coding genes, far fewer than the originally predicted number of 80,000–100,000 genes.
- The average size of a human gene is ~25 kb, including gene-regulatory regions, introns, and exons. On average, mRNAs produced by human genes are ~3000 nt long.
- Many human genes produce more than one protein through alternative splicing, thus enabling human cells to produce a much larger number of proteins (perhaps as many as 200,000) from only ~20,000 genes.
- More than 50 percent of human genes show a high degree of sequence similarity to genes in other organisms; however, more than 40 percent of the genes identified have no known molecular function.
- Genes are not uniformly distributed on the 24 human chromosomes. Gene-rich clusters are separated by gene-poor "deserts" that account for 20 percent of the genome. These deserts correlate with G bands seen in stained chromosomes. Chromosome 19 has the highest gene density, and chromosome 13 and the Y chromosome have the lowest gene densities.
- Chromosome 1 contains the largest number of genes, and the Y chromosome contains the smallest number.
- Human genes are larger and contain more and larger introns than genes in the genomes of invertebrates, such as *Drosophila*. The largest known human gene encodes dystrophin, a muscle protein. This gene, associated in mutant form with muscular dystrophy, is 2.5 Mb in length (Chapter 14), larger than many bacterial chromosomes. Most of this gene is composed of introns.
- The number of introns in human genes ranges from 0 (in histone genes) to 234 (in the gene for *titin*, which encodes a muscle protein).

that there are only around 20,000 protein-coding genes. Recall that the number of genes had originally been estimated to be about 100,000, based in part on a prediction that human cells produce about 100,000 proteins. At least half of the genes show sequence similarity to genes shared by many other organisms, and as you will learn in Section 21.7, a majority of human genes are similar in sequence to genes from closely related species such as chimpanzees. There is still no consensus among scientists worldwide about the exact number of human genes. One reason is that it is unclear whether or not many of the presumed genes produce functional proteins. Currently, annotation predicts that the human genome encodes approximately 21,000 proteins. But (see Table 21.1), due to posttranslational modifications and other processes, the total number of proteins present in human cells is thought to be anywhere from approximately 200,000 to 1 million! Genome scientists continue to annotate the genome, and as mentioned earlier, functional genomics studies have important roles in determining whether or not computational predictions about the number of protein-coding and non—protein-coding genes are accurate.

The number of genes is much lower than the number of predicted proteins in part because many genes code for multiple proteins through **alternative splicing.** Recall from earlier in the text (see Chapter 13), that alternative splicing patterns can generate multiple mRNA molecules, and thus multiple proteins, from a single gene, through different combinations of intron—exon splicing arrangements. Initial estimates suggested that over 50 percent of human genes undergo alternative splicing to produce multiple transcripts and multiple proteins. Recent studies suggest that ~94–95 percent of human pre-mRNAs contain multiple exons that are processed to produce multiple transcripts and potentially multiple different protein products. Clearly, alternative splicing produces an incredible diversity of proteins beyond simple predictions based on the number of genes in the human genome.

During the HGP, functional categories were assigned for human genes, primarily on the basis of

1. Functions determined previously (for example, from recombinant DNA cloning of human genes and known mutations involved in human diseases)
2. Comparisons to known genes and predicted protein sequences from other species
3. Predictions based on annotation and analysis of protein functional domains and motifs

Although functional categories and assignments continue to be revised, at the time the HGP was completed, the functions of over 40 percent of human genes were unknown. Determining human gene functions, deciphering complexities of gene-expression regulation and gene interaction, and uncovering the relationships between human genes and phenotypes are among the many ongoing challenges for genome scientists.

Individual Variations in the Human Genome

The HGP originally revealed that in all humans, regardless of racial and ethnic origins, the genomic sequence is approximately 99.9 percent the same. As we discuss in other chapters, most genetic differences between humans result from **single-nucleotide polymorphisms (SNPs)** and **copy number variations (CNVs).** Recall (from Chapter 5) that SNPs are single-base changes in the genome and variations of many SNPs are associated with disease conditions. For example, SNPs cause sickle-cell anemia and cystic fibrosis. Later in the text (see Chapter 22), we will examine how SNPs can be detected and used for diagnosis and treatment of disease.

After the draft sequence of the human genome was completed, it initially appeared that most genetic variations between individuals (the 0.1 percent differences) were due to SNPs. While SNPs are important contributing factors to genome variation, structural differences that we discussed earlier in the text (see Chapter 12) such as deletions, duplications, inversions, and CNVs, which can span millions of base pairs of DNA, play much more important roles in genome variation than previously thought. Recall that CNVs are duplications or deletions of relatively large sections of DNA on the order of several hundred or several thousand base pairs (see Chapter 8). Many of the CNVs that vary the most among genomes appear to be at least 1 kilobase.

Although most human DNA is present in two copies per cell, one from each parent, CNVs are segments of DNA that are duplicated or deleted, resulting in variations in the number of copies of a DNA segment inherited by individuals. In some cases CNVs are major deletions at the exon level or involving entire genes; other deletions affect gene function by frameshifts in the reading code. CNVs that are duplicated can result in overexpression of a particular gene, yet many deleted and duplicated CNVs do not present clearly identifiable phenotypes.

Current estimates of the number of CNVs in an individual genome range from about 12 to perhaps 4—5 dozen per person. Some studies estimate that there may be as many as 1500 CNVs greater than 1 kb among the human genome. Other studies claim there are more than 1.5 million deletions of less than 100 bp that contribute to genome variation between individuals.

Accessing the Human Genome Project on the Internet

It is now possible to access databases and other sites on the Internet that display maps for all human chromosomes. You will visit a number of these databases in Exploring Genomics exercises. **Figure 21.9(a)** displays a partial gene map for chromosome 12 that was taken from an NCBI database called Map Viewer. You may already have used Map Viewer for the Exploring Genomics exercises earlier in the text (see Chapters 5 and 12). The first image shows an ideogram, or

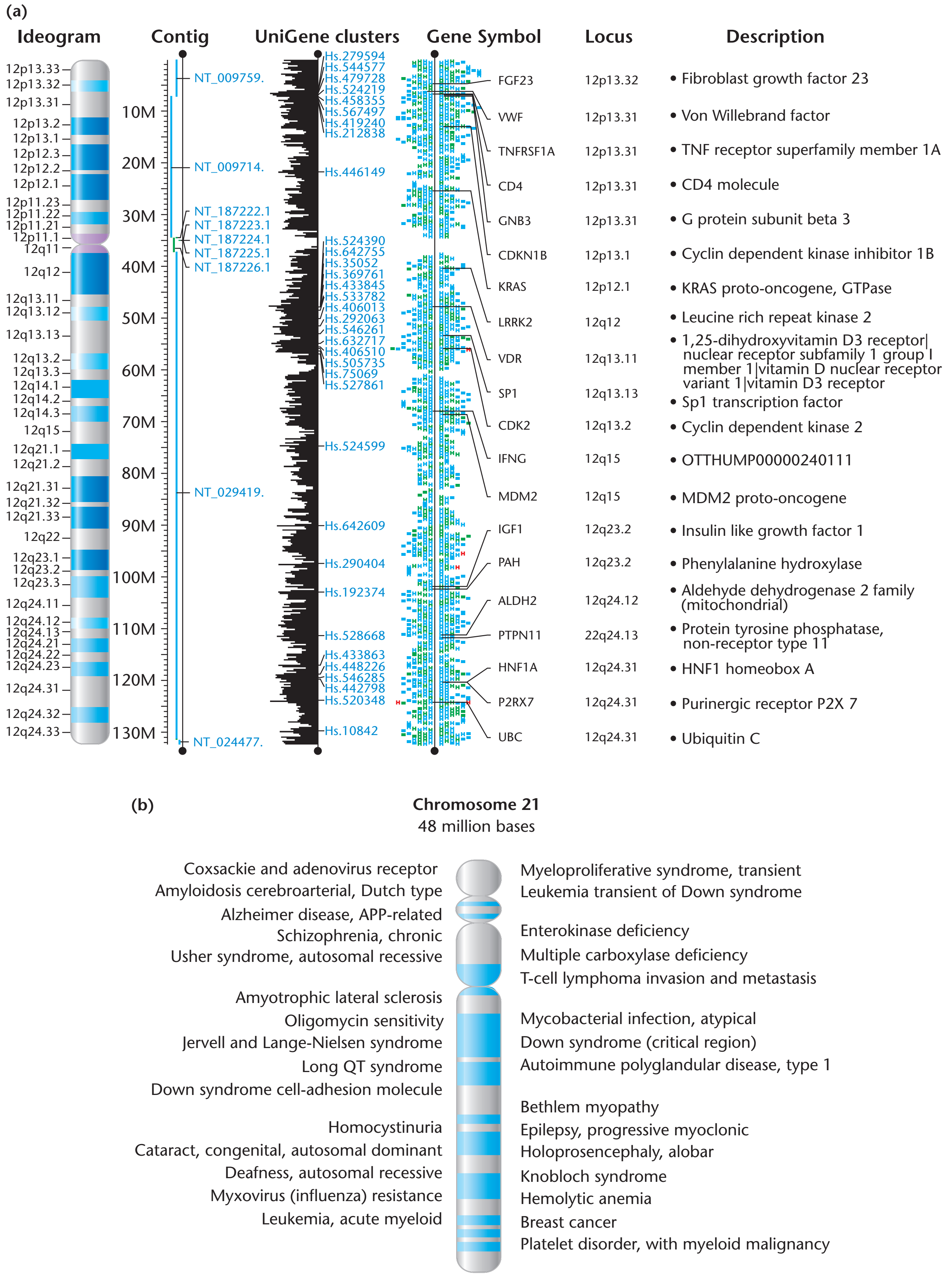

(a)
Ideogram
Contig
UniGene clusters
Gene Symbol
Locus
Description
12p13.33
12p13.32
12p13.31
12p13.2
12p13.1
12p12.3
12p12.2
12p12.1
12p11.23
12p11.22
12p11.21
12p11.1
12q11
12q12
12q13.11
12q13.12
12q13.13
12q13.2
12q13.3
12q14.1
12q14.2
12q14.3
12q15
12q21.1
12q21.2
12q21.31
12q21.32
12q21.33
12q22
12q23.1
12q23.2
12q23.3
12q24.11
12q24.12
12q24.13
12q24.21
12q24.22
12q24.23
12q24.31
12q24.32
12q24.33
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
NT_009759.
NT_009714.
NT_187222.1
NT_187223.1
NT_187224.1
NT_187225.1
NT_187226.1
NT_029419.
NT_024477.
Hs.279594
Hs.544577
Hs.479728
Hs.524219
Hs.458355
Hs.567497
Hs.419240
Hs.212838
Hs.446149
Hs.524390
Hs.642755
Hs.35052
Hs.369761
Hs.433845
Hs.533782
Hs.406013
Hs.292063
Hs.546261
Hs.632717
Hs.406510
Hs.505735
Hs.75069
Hs.527861
Hs.524599
Hs.642609
Hs.290404
Hs.192374
Hs.528668
Hs.433863
Hs.448226
Hs.546285
Hs.442798
Hs.520348
Hs.10842
FGF23 12p13.32 • Fibroblast growth factor 23
VWF 12p13.31 • Von Willebrand factor
TNFRSF1A 12p13.31 • TNF receptor superfamily member 1A
CD4 12p13.31 • CD4 molecule
GNB3 12p13.31 • G protein subunit beta 3
CDKN1B 12p13.1 • Cyclin dependent kinase inhibitor 1B
KRAS 12p12.1 • KRAS proto-oncogene, GTPase
LRRK2 12q12 • Leucine rich repeat kinase 2
VDR 12q13.11 • 1,25-dihydroxyvitamin D3 receptor| nuclear receptor subfamily 1 group I member 1|vitamin D nuclear receptor variant 1|vitamin D3 receptor
SP1 12q13.13 • Sp1 transcription factor
CDK2 12q13.2 • Cyclin dependent kinase 2
IFNG 12q15 • OTTHUMP00000240111
MDM2 12q15 • MDM2 proto-oncogene
IGF1 12q23.2 • Insulin like growth factor 1
PAH 12q23.2 • Phenylalanine hydroxylase
ALDH2 12q24.12 • Aldehyde dehydrogenase 2 family (mitochondrial)
PTPN11 22q24.13 • Protein tyrosine phosphatase, non-receptor type 11
HNF1A 12q24.31 • HNF1 homeobox A
P2RX7 12q24.31 • Purinergic receptor P2X 7
UBC 12q24.31 • Ubiquitin C
(b)
Chromosome 21
48 million bases
Coxsackie and adenovirus receptor
Amyloidosis cerebroarterial, Dutch type
Alzheimer disease, APP-related
Schizophrenia, chronic
Usher syndrome, autosomal recessive
Amyotrophic lateral sclerosis
Oligomycin sensitivity
Jervell and Lange-Nielsen syndrome
Long QT syndrome
Down syndrome cell-adhesion molecule
Homocystinuria
Cataract, congenital, autosomal dominant
Deafness, autosomal recessive
Myxovirus (influenza) resistance
Leukemia, acute myeloid
Myeloproliferative syndrome, transient
Leukemia transient of Down syndrome
Enterokinase deficiency
Multiple carboxylase deficiency
T-cell lymphoma invasion and metastasis
Mycobacterial infection, atypical
Down syndrome (critical region)
Autoimmune polyglandular disease, type 1
Bethlem myopathy
Epilepsy, progressive myoclonic
Holoprosencephaly, alobar
Knobloch syndrome
Hemolytic anemia
Breast cancer
Platelet disorder, with myeloid malignancy

FIGURE 21.9 (a) A gene map for chromosome 12 from the NCBI database Map Viewer. (b) Partial map of disease genes on human chromosome 21. Maps such as this depict genes thought to be involved in human genetic disease conditions.

cytogenetic map, of chromosome 12. To the right of the ideogram is a column showing the contigs (arranged vertically) that were aligned to sequence this chromosome. The UniGene column displays a histogram representation of gene density on chromosome 12. Notice that relatively few genes are located near the centromere. Finally, gene symbols, loci, and gene names (by description) are provided for selected genes; in this figure only 20 genes are identified. When accessing these maps on the Internet, one can magnify, or zoom in on, each region of the chromosome, revealing all genes mapped to a particular area.

You can see that most of the genes listed in Figure 21.9(a) have been assigned descriptions based on the functions of their products: Some are transmembrane proteins; some are enzymes such as kinases; some are receptors, including several involved in olfaction; and so on.

The HGP's most valuable contribution will perhaps be the identification of disease genes and the development of new treatment strategies as a result. Thus, extensive maps have been developed for genes implicated in human disease conditions. The disease gene map of chromosome 21 shown in **Figure 21.9(b)** indicates genes involved in amyotrophic lateral sclerosis (ALS), Alzheimer disease, cataracts, deafness, and several different cancers. In a later chapter (see Chapter 22), we discuss implications of the HGP for the identification of genes involved in human genetic diseases, and for disease diagnosis, detection, and gene therapy applications.

Another excellent resource on the Internet is the PANTHER (Protein ANalysis THrough Evolutionary Relationships) database. This provides functional classifications of genes and encoded proteins using data from functional genomics, experimental approaches, and other analyses. Refer to PDQ 27 for an exercise that will take you to a pie chart view of current functional categories for human genes.

21.6 The "Omics" Revolution Has Created a New Era of Biological Research

The Human Genome Project and the development of genomics techniques have been largely responsible for launching a new era of biological research—the era of "omics." It seems that every year, more areas of biological research are being described as having an omics connection. Some examples of "omics" are:

- Proteomics—the analysis of all the proteins in a cell or tissue
- Metabolomics—the analysis of proteins and enzymatic pathways involved in cell metabolism
- Glycomics—the analysis of the carbohydrates of a cell or tissue
- Toxicogenomics—the analysis of the effects of toxic chemicals on genes, including mutations created by toxins and changes in gene expression caused by toxins
- Metagenomics—the analysis of genomes of organisms collected from the environment
- Pharmacogenomics—the development of customized medicine based on a person's genetic profile for a particular condition
- Transcriptomics—the analysis of all expressed genes in a cell or tissue

We will consider several of these genomics disciplines in other parts of this chapter.

After the HGP, What's Next?

Since completion of a reference sequence of the human genome, studies have continued at a very rapid pace. For example, as a result of the HGP, many other major theme areas for human genome research have emerged, including cancer genome projects, analysis of the epigenome (including a Human Epigenome Project that is creating hundreds of maps of epigenetic changes in different cell and tissue types and evaluating potential roles of epigenetics in complex diseases), and characterization of SNPs (the International HapMap Project) and CNVs for their role in genome variation, disease, and pharmacogenomics applications. We will discuss aspects of a cancer genome project (Cancer Genome Atlas Project) later in the text (see Chapter 24). The epigenome was covered in depth in Chapter 19. SNPs and pharmacogenomics are discussed later as well (see Special Topic Chapter 3—Genomics and Precision Medicine). Here we consider several examples of genome research that are extensions of the HGP.

Personal Genome Projects

As we discussed earlier in this chapter (and in Chapter 20), next- and third-generation sequencing technologies, capable of generating sequence reads at higher speeds with greater accuracy, have greatly reduced the cost of DNA sequencing, and expectations for continued cost reductions along with continued technological advances are high (see **Figure 21.10**). These expectations have led several companies to propose WGS for individual people—a **personal genomics** approach. Two programs funded by the National Institutes of Health challenged scientists to develop sequencing technologies to complete a human genome for $1000 by 2014 (see the Genetics, Technology, and Society essay in Chapter 22).

As you learned earlier in the text (see Chapter 20), several companies have now developed technology that can sequence a genome for less than $1000. Whether the

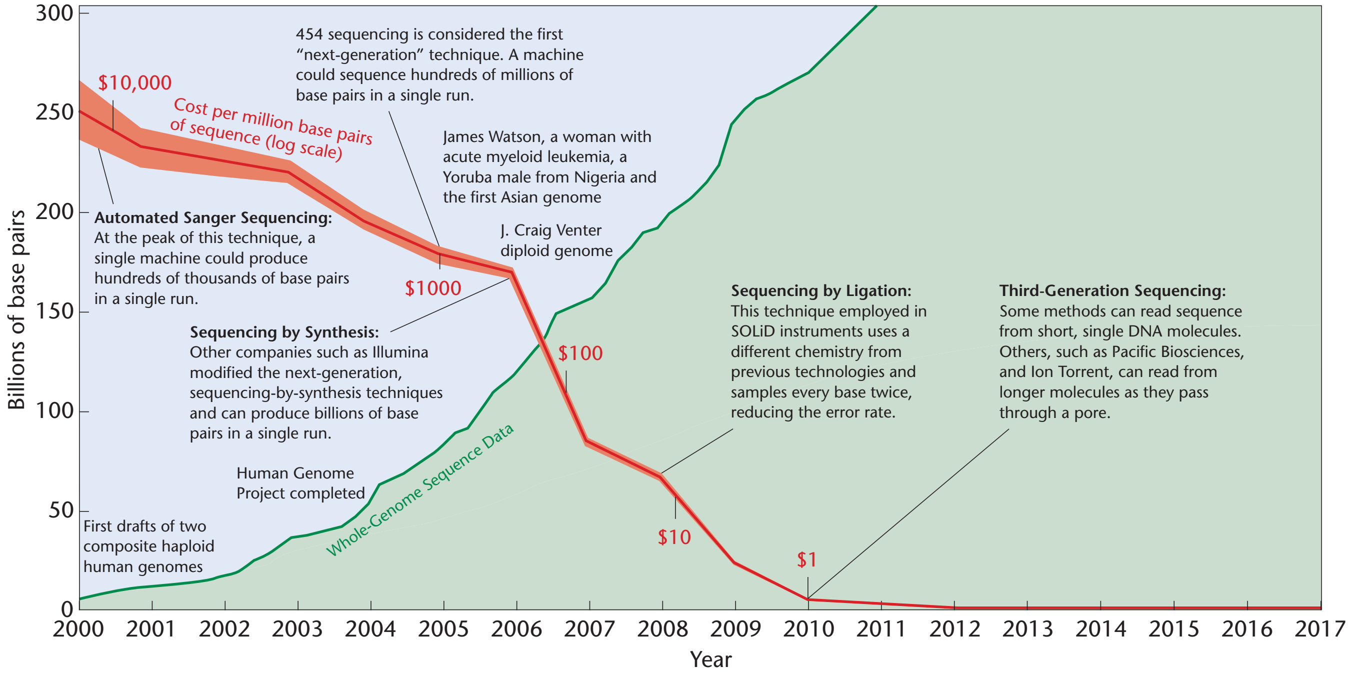


FIGURE 21.10 Human genome sequence explosion. Sequencing costs have steadily declined since 2000 due to innovations in sequencing technology. As a result, notice that the number of individual genomes sequenced has dramatically increased.

$1000 mark represents the costs of reagents to sequence a genome or actual costs when sequence preparation, labor, and analysis of the genome are taken into account can be debated.

As you will learn later in in the text (see Chapter 22), having somebody such as a geneticist analyze your personal genome data and consider how genome variations may affect your health may be expensive. So even if the cost of sequencing a genome is less than $1000, interpreting genome data to make sense for medical treatment may cost thousands of dollars more. Regardless of how the actual cost of sequencing a genome is calculated, the modern cost is substantially lower than $3 billion.

As of late 2016, an estimated 300,000 individual human genomes have been sequenced. One especially intriguing example of personal genomics involves a genetics researcher who examined his own genome for insight about a medical condition. Dr. Richard Gibbs of Baylor College of Medicine led a group that sequenced the whole genome of his colleague Dr. James Lupski, a medical geneticist who has **Charcot-Marie-Tooth (CMT) disease.** This disease is a neurological condition that causes muscle weakness in the extremities. Interestingly, mutations in over 30 genes, many of which were identified by Dr. Lupski, are involved in CMT, although Lupski did not carry any of these mutations. A comparison of Lupski's genome to the HGP reference sequence revealed many SNPs and other variations, but it was unclear how many of these were simply sequencing errors or variations that were not involved in CMT.

Focusing on genes previously linked to CMT and other neurological conditions, researchers found that Lupski's genome had two different mutations in the gene *SH3TC2,* which is expressed in Schwann cells that wrap around certain neurons to form the myelin sheath essential for impulse conductions in nerves. In one *SH3TC2* allele a nonsense mutation was revealed, and in Lupski's second allele for *SH3TC2* a new missense mutation was found. When genetic tests for these alleles were carried out on Lupski's parents and seven siblings, the nonsense mutation was found in one parent and two siblings who did not have the disease. The missense mutation was found in another parent and one grandparent, neither of whom had the disorder. Only siblings who inherited both mutated alleles had CMT disease. Many consider this the first clinically relevant success for personal genome sequencing—at least for identifying disease genes.

Somatic Genome Mosaicism and the Emerging Pangenome

The HGP pooled samples from multiple individuals to create a *haploid* reference genome. In contrast, personal genome projects (PGPs) sequence a *diploid* genome. Personal genome projects generate sequences of millions of short DNA fragments from *maternal and paternal chromosomes* that are mapped onto the reference genome. Such projects indicate that haploid reference genome comparisons often underestimate the extent of genome variation between individuals by five-fold or more. For example, when Craig Venter's personal genome was analyzed, over

4 million variations were found between his maternal and paternal chromosomes alone.

From what we are learning about personal genomes, genome variation between individuals may be closer to 0.5 percent than to the 0.1 percent predicted by the HGP, and in a 3-billion-bp genome this is a significant difference in sequence variation. Integrating genome data from several complete personal genomes of individuals from different ethnic groups will also be of great value in evolutionary genetics to address fundamental questions about human diversity, ancestry, and migration patterns.

Personal genome projects are revealing that there can be significant **genome mosaicism** in human somatic cells. It is now apparent that cells in an individual person do not all contain identical genomes. We have to think of an individual as being made up of a population of cells, each with its own unique personal genome. Somatic mosaicism can result from errors in DNA replication, creating aneuploidy, CNVs, SNPs, and other variations that accumulate as cells divide during development. In somatic cells these variations are passed to daughter cells during mitosis, but typically not transmitted to offspring.

In the past, finding mosaicism was not easy because researchers typically pooled DNA from *large numbers of cells* (such as blood cells), sequenced samples multiple times, then used software to develop a likely reference sequence. But now, because of the sensitivity of modern sequencing methods, including the ability to sequence *individual cells* (which we will discuss further in Chapter 22), mosaicisms for SNPs and CNVs have been found in skin, brain, blood, and stem cells from the same individual, among other cells types. Estimates suggest that over a third of skin fibroblasts have CNVs that contribute to mosaicism in humans.

We are only beginning to understand the frequency and effects of genetic mosaicism on health and disease. For example, Lupski's group at Baylor has also identified CNVs in patients with rare disorders. One such example is **Smith–Magenis syndrome (SMS)**, a developmental disorder linked to microdeletions on the p-arm of chromosome 17. SMS affects approximately 1 in 25,000 individuals and results in cognitive disabilities, difficulty sleeping, and physical features such as a broad face. Microdeletions on the p-arm of chromosome 17 and mosaicisms for these deletions affect the recurrence of SMS (**Figure 21.11**).

In another study, scientists from the Scripps Research Institute analyzed neurons harvested from postmortem brains. They found that the *APP* gene, encoding amyloid precursor protein, was mosaically amplified in brains from Alzheimer patients with some neurons containing up to 12 copies of the *APP* gene.

Newer sequencing methods and expansion of personal genome projects are demonstrating significant **genomic variation** not just in humans but in most species. It has now become apparent that genomic variation among individuals is much more prevalent than can be determined by a reference sequence. In bacteria such as *Streptococcus*, for example, several dozen different genes can exist between isolates of the same strain. Genome scientists in many fields have now replaced single reference genomes with the concept of the **pangenome** to describe all distinct genes and variations in a species (**Figure 21.12**).

Whole-Exome Sequencing

While we have thus far focused on sequencing the entire genome, it is worthwhile to recall that only about 2 percent of the genome sequence consists of protein-coding genes. Thus, in personal genomic analysis, the focus has shifted toward **whole-exome sequencing (WES),** that is, sequencing only the 180,000 exons in a person's genome. WES reveals mutations involved in disease by focusing only on exons as protein-coding segments of the genome, thus there are more disease-related genetic variations in the exome than in other regions of the genome. WES can be done at a cost of much less than $1000 with greater than 100 × coverage

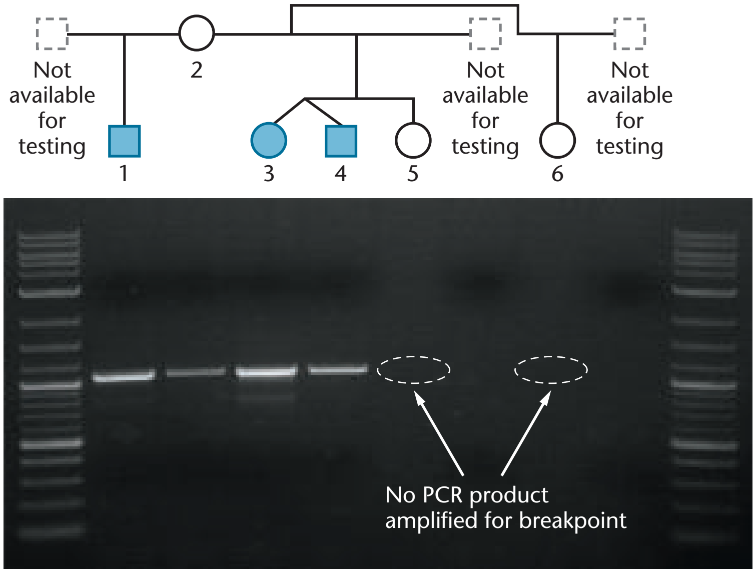

Not available for testing
2
Not available for testing
Not available for testing
1
3
4
5
6
No PCR product amplified for breakpoint

FIGURE 21.11 Smith–Magenis syndrome (SMS) recurrence demonstrates mosaicism. Shown here are results from blood tests for six siblings. PCR was used to amplify a breakpoint junction for the microdeletion on chromosome 17. Notice that children 1, 3, and 4 (black markers) have the disease. Children 5 and 6 do not have the disease nor do they have the breakpoint. Child 2 does not have SMS but is still a carrier for the deletion due to mosaicism. Notice how the amplified segment of DNA for child 2 shows up as a less intense (less bright) band on the gel compared to that for children l, 3, and 4. This demonstrates mosaicism for the deletion, which can be confirmed by sequencing.

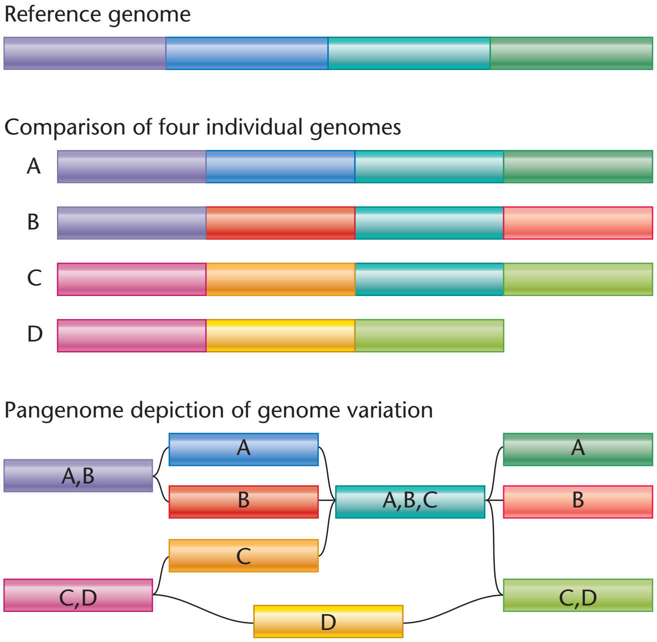

Reference genome
Comparison of four individual genomes
A
B
C
D
Pangenome depiction of genome variation
A,B
A
B
C
C,D
D
A,B,C
A
B
C,D

FIGURE 21.12 A pangenome attempts to visualize all genomic segments and gene variations found in a species. Notice that there are variations in individual genomes not represented in the reference genome, but these variations are included in the pangenome.

(coverage is the percentage of bases in a DNA sequence that have been sequenced multiple times). Of course, a limitation of this approach is its failure to identify mutations in gene-regulatory regions that influence gene expression. As the cost of WGS continues to drop, scientists and clinicians trying to detect disease-causing mutations are debating whether it makes sense to simply sequence the entire genome or to just sequence exomes first to find mutations.

In 2015, after seven years of work, a group of scientists called the 1000 Genomes Project Consortium reported on the genomes of 2504 individuals from 26 populations representing Europe, East Asia, South Asia, Africa, and the Americas. WGS and WES data revealed nearly 88 million bi- and multi-allelic SNPs and many other structural variations such as CNVs. One interpretation of this work is that it reveals clear variations in individuals and associates particular diseases with geographic or ancestral background. Thus sequencing genomes of individuals from diverse populations can help us better understand the spectrum of human genetic variation and to learn the causes of genetic diseases across diverse groups. We will come back to the topic of WES later in the text when we discuss genetic testing (see Chapter 22).

Encyclopedia of DNA Elements (ENCODE) Project

In 2003, a few months after the announcement that the human genome had been sequenced, a group of about three dozen research teams around the world began the **Encyclopedia of DNA elements (ENCODE) project**. Using both experimental and computational approaches, a main goal of ENCODE was to identify and analyze all functional elements of the genome, including those that regulate the expression of human genes (such as transcriptional start sites, promoters, and enhancers). ENCODE projects have also been initiated for mouse, worm, and fly genomes.

Because only a relatively small percentage (less than 2 percent) of the human genome codes for proteins, ENCODE focused heavily not on genes but on all the rest of the sequences, commonly referred to as "junk" DNA. So what are all these other bases in the genome doing? The term *junk DNA* has always been a misnomer. We know that such sequences are important for chromosome structure, the regulation of gene expression, and other roles. Just because these sequences themselves do not code for protein does not mean that they are unimportant. Non–protein-coding sequences were discussed in greater detail earlier in the text (see Chapter 18).

ENCODE studied gene expression in 147 different cell types because genome activity differs from cell to cell. After about a decade of research and a cost of $288 million, in 2012 a group of 30 research papers were published revealing the major initial findings of the ENCODE project.

Selected highlights of what ENCODE revealed in those initial papers and recent highlights include the following:

- The *majority*, ~80 percent, of the human genome is considered functional. This is partly because large segments of the genome are transcribed into RNA. Most of these RNAs do not encode proteins. These various RNAs include tRNA, rRNAs, and miRNAs, and long noncoding RNAs (lncRNAs)—defined as non-protein-coding transcripts longer than 200 nucleotides. A conservative estimate is that there may be over 17,000 genes for lncRNAs. It may turn out that the number of noncoding RNA sequences will outnumber protein-coding genes.
- The functional sequences also include gene-regulatory regions: ~70,000 promoter regions and nearly 400,000 enhancer regions.
- There are 20,687 protein-coding genes in the human genome.
- A total of 11,224 sequences are characterized as pseudogenes, previously thought to be inactive in all individuals. Some of these are inactive in most individuals but occasionally active in certain cell types of some individuals, which may eventually warrant their reclassification as active, transcribed genes and not pseudogenes.
- SNPs associated with disease are enriched within noncoding functional elements of the genome, often residing near protein-coding genes.

The ENCODE findings have broadly defined the functional roles of the genome to include encoding proteins or noncoding RNAs and displaying biochemical properties such as binding regulatory proteins that influence transcription or chromatin structure. It is worth noting, however, that a relatively large body of geneticists and other scientists do not agree with ENCODE's definition of functional sequences. One reason cited is that ENCODE did not adequately address many of the repetitive sequences in the genome such as transposons, LINEs, SINEs, and other sequences such as telomeres and centromeres (see Chapter 12). There has also been significant debate about the value of ENCODE, given the cost of the project. But research teams are using information from ENCODE to identify risk factors for certain diseases, with the hopes of developing appropriate cures and treatments.

Nutrigenomics Considers Genetics and Diet

As evidence of the impact of genomics, a field of nutritional science called nutritional genomics, or **nutrigenomics,** has emerged. Nutrigenomics focuses on understanding the interactions between diet and genes. We have all had routine medical tests for blood pressure, blood sugar levels, and heart rate. Based on these tests, your physician may recommend that you change your diet and exercise more to lose weight, or that you reduce your intake of sodium to help lower your blood pressure.

Now several companies claim to provide nutrigenomics tests that analyze your genomes for genes thought to be associated with different medical conditions linked to nutrient metabolism. The companies then provide a customized nutrition report, recommending diet changes for improving your health and preventing illness, based on your genes! It is important to note that these tests have not yet been validated as accurate and they have not been approved by the U.S. Food and Drug Administration. It remains to be seen whether this approach as currently practiced is of valid scientific or nutritional value.

No Genome Left Behind and the Genome 10K Plan

Without question, new sequencing technologies that have been developed are an important part of the transformational effect the HGP has had on modern biology.

Recent headline-grabbing genomes that have been completed include:

- **Apple and tomato.** The apple, which has more than 57,000 genes, and the tomato, which has 31,760 genes, each have more genes than humans!
- **Potato.** This plant, which shares 92 percent of its DNA with tomatoes, turns out to be a fruit.
- **Chickpea.** It is one of the earliest cultivated legumes and the second most widely grown legume after the soybean.
- **Red-spotted newt.** This newt has a genome of almost 10 times the size of a human genome.

Modern sequencing technologies are asking some to consider the question, "What would you do if you could sequence everything?" In 2009, partners around the world, including genome scientists and zoo and museum curators, began work on sequencing 10,000 vertebrate genomes—the **Genome 10K project.** Shortly after the HGP finished, the National Human Genome Research Institute (NHGRI) assembled a list of mammals and other vertebrates as priorities for genome sequencing in part because of their potential benefit for learning about the human genome through comparative genomics. Genome 10K will also provide insight into genome evolution and speciation.

Stone-Age Genomics

In yet another example of how genomics has taken over areas of DNA analysis, a number of labs around the world are involved in analyzing "ancient" DNA. These so-called **stone-age genomics** studies are generating fascinating data from miniscule amounts of ancient DNA obtained from bone and other tissues such as hair that are tens of thousands to about 700,000 years old, and often involve samples from extinct species. Analysis of DNA from a 2400-year-old Egyptian mummy, bison, mosses, platypus, mammoths, Pleistocene-age cave bears and polar bears, coelacanths, and Neanderthals are some of the most prominent examples of stone-age genomics. In 2013, scientists reported the oldest intact genome sequence to be successfully analyzed to date. It came from a 700,000-year-old bone fragment from an ancient horse uncovered from the frozen ground in the Yukon Territory of Canada. This result is interesting in part because evolutionary biologists have used genomic data to estimate that ancient ancestors of modern horses branched off from other animal lineages around 4 million years ago—about twice as long ago as prior estimates.

A little over a decade ago, researchers published about 13 million bp of a sequence from a 27,000-year-old woolly mammoth found frozen and nearly intact in Siberia. This study revealed a ~98.5 percent sequence identity between mammoths and African elephants. Subsequent studies by other scientists have used whole-genome shotgun sequencing of mitochondrial and nuclear DNA from

Siberian mammoths to provide data on the mammoth genome. These studies suggest that the mammoth genome differs from the African elephant by as little as 0.6 percent. These studies are also great demonstrations of how stable DNA can be under the right conditions, particularly when frozen.

In Section 21.7 we discuss recent work on the Neanderthal genome. Obtaining the genome of a human ancestor this old was previously unimaginable. This work is providing new insights into our understanding of human evolution.

21.7 Comparative Genomics Analyzes and Compares Genomes from Different Organisms

As of 2016, over 21,000 whole genomes have been sequenced—including many model organisms and a number of viruses. This is quite extraordinary progress in a relatively short time span! Among these organisms are yeast (*S. cerevisiae*, the first eukaryotic genome to be completely sequenced), bacteria such as *E. coli*, the nematode roundworm (*Caenorhabditis elegans*), the thale cress plant (*Arabidopsis thaliana*), mice (*M. musculus*), zebrafish (*Danio rerio*), and of course *Drosophila*. In recent years, genomes of chimpanzees, dogs, chickens, gorillas, sea urchins, honey bees, pigs, pufferfish, rice, and wheat have all been sequenced.

These studies have demonstrated not only significant differences in genome organization between bacteria and eukaryotes but also many similarities between genomes of nearly all species. Analysis of the growing number of genome sequences confirms that all living organisms are related and descended from a common ancestor. Similar gene sets are used by organisms for basic cellular functions, such as DNA replication, transcription, and translation. These genetic relationships are the rationale for using model organisms to study inherited human disorders; the effects of the environment on genes; and interactions of genes in complex diseases, such as cardiovascular disease, diabetes, neurodegenerative conditions, and behavioral disorders.

Comparative genomics compares the genomes of different organisms to answer questions about genetics and other aspects of biology. It is a field with many research and practical applications, including gene discovery and the development of model organisms to study human diseases. It also incorporates the study of gene and genome evolution and the relationship between organisms and their environment. Comparative genomics uses a wide range of techniques and resources, such as the construction and use of nucleotide and protein databases containing nucleic acid and amino acid sequences, fluorescent *in situ* hybridization (FISH), and the creation of gene knockout animals. Comparative genomics can reveal genetic differences and similarities between organisms to provide insight into how those differences contribute to differences in phenotype, life cycle, or other attributes, and to ascertain the evolutionary history of those genetic differences.

Bacterial and Eukaryotic Genomes Display Common Structural and Functional Features and Important Differences

Since most bacteria have small genomes amenable to shotgun cloning and sequencing, many early genome projects focused on bacteria, and more than 1000 additional projects to sequence bacterial genomes are now under way. Many of the bacterial genomes already sequenced are from organisms that cause human diseases, such as cholera, tuberculosis, and leprosy. Traditionally, the bacterial genome was thought of as relatively small (less than 5 Mb) and contained within a single circular DNA molecule. *E. coli*, used as the prototypical bacterial model organism in genetics, has a genome with these characteristics. However, the flood of genomic information now available has challenged the validity of this viewpoint for bacteria in general. Although most bacterial genomes are small, their sizes vary across a surprisingly wide range. In fact, there is some overlap in size between larger bacterial genomes (30 Mb in *Bacillus megaterium*) and smaller eukaryotic genomes (12.1 Mb in yeast).

Gene number in bacterial genomes also demonstrates a wide range, from less than 500 to more than 5000 genes, a ten-fold difference. A bacterial genome with one of the largest number of genes discovered so far belongs to the cyanobacterium *Prochlorococcus*, the smallest but most abundant photosynthetically-active cell in the ocean that has been discovered to date. Different strains of *Prochloroccocus* have an estimated 80,000 genes. Some estimate that this tiny microbe accounts for ~5% of global photosynthesis!

In addition, although many bacteria have a single, circular chromosome, there is substantial variation in chromosome organization and number among bacterial species. An increasing number of genomes composed of linear DNA molecules are being identified, including the genome of *Borrelia burgdorferi*, the organism that causes Lyme disease. Sequencing of the *Vibrio cholerae* genome (the organism responsible for cholera) revealed the presence of two circular chromosomes. Other bacteria that have genomes with two or more chromosomes include *Rhizobium radiobacter* (formerly *Agrobacterium tumefaciens*), *Deinococcus radiodurans*, and *Rhodobacter sphaeroides*. The finding that some bacterial species have multiple chromosomes raises questions both about how replication and segregation of their chromosomes are coordinated during cell division and about what undiscovered mechanisms of gene regulation may exist in bacteria. The answers may provide clues about the evolution of multichromosome eukaryotic genomes.

We can make two generalizations about the organization of protein-coding genes in bacteria. First, gene density is very high, averaging about one gene per kilobase of DNA. For example, the genome of *E. coli* strain K12, which was the second bacterial genome to be sequenced, is 4.6 Mb in size, and it contains 4289 protein-coding genes in its single, circular chromosome. This close packing of genes in bacterial genomes means that a very high proportion of the DNA (approximately 85 to 90 percent) serves as coding DNA. Typically, only a small amount of a bacterial genome is noncoding DNA, often in the form of regulatory sequences or of transposable elements that can move from one place to another in the genome.

The second generalization we can make is that bacterial genomes contain operons. Recall from earlier in the text (see Chapter 16) that operons contain multiple genes functioning as a transcriptional unit whose protein products are part of a common biochemical pathway. In *E. coli*, 27 percent of all genes are contained in operons (almost 600 operons).

The basic features of eukaryotic genomes are similar in different species, although genome size in eukaryotes is highly variable (**Table 21.2**). Genome sizes range from about 10 Mb in fungi to over 100,000 Mb in some flowering plants (a ten thousand-fold range); the number of chromosomes per genome ranges from two to the hundreds (about a hundred-fold range), but the number of genes varies much less dramatically than either genome size or chromosome number.

Eukaryotic genomes have several features not found in bacteria:

- **Gene density.** Whereas in bacteria, gene density is generally close to 1 gene per kilobase, eukaryotic genomes exhibit a wide range of gene density. In yeast, there is about 1 gene/2 kb; in *Drosophila*, there is about 1 gene/13 kb; and in humans, gene density varies greatly from chromosome to chromosome. Human chromosome 22 has about 1 gene/64 kb, while chromosome 13 has 1 gene/155 kb of DNA.
- **Introns.** Most eukaryotic genes contain introns. There is wide variation among genomes in the number of introns they contain and also wide variation from gene to gene. The entire yeast genome has only 239 introns, whereas just a single gene in the human genome can contain more than 100 introns. Regarding intron size, generally the size in eukaryotes is correlated with genome size. Smaller genomes have smaller average intron sizes, and larger genomes have larger average intron sizes. But there are exceptions. For example, the genome of the pufferfish (*Takifugu rubripes*) has relatively few introns.

TABLE 21.2 **Comparison of Selected Genomes**

Organism (Scientific Name)	Approximate Size of Genome [in million (megabase, Mb) or billion (gigabase, Gb) base pairs] (Date Completed)	Diploid (2*n*) Chromosome Number	Number of Genes	Approximate Percentage of Genes Shared with Humans
African clawed frog (*Xenopus laevis*)	3.1 Gb (2016)	36	~45,000	70%
Bacterium (*Escherichia coli*)	4.6 Mb (1997)	1	4403	Not determined
Chicken (*Gallus gallus*)	1 Gb (2004)	78	~20,000–23,000	60%
Dog (*Canis familiaris*)	2.5 Gb (2003)	78	~18,400	75%
Chimpanzee (*Pan troglodytes*)	~3 Gb (2005)	48	~20,000–24,000	98%
Fruit fly (*Drosophila melanogaster*)	165 Mb (2000)	8	~13,600	50%
Human (*Homo sapiens*)	3.1 Gb (2004)	46	~20,000	100%
Mouse (*Mus musculus*)	~2.5 Gb (2002)	40	~30,000	80%
Pig (*Sus scrofa*)	~3 Gb (2012)	38	21,640	84%
Rat (*Rattus norvegicus*)	~2.75 Gb (2004)	42	~22,000	80%
Rhesus macaque (*Macaca mulatta*)	2.87 Gb (2007)	42	~20,000	93%
Rice (*Oryza sativa*)	389 Mb (2005)	24	~41,000	Not determined
Roundworm (*Caenorhabditis elegans*)	97 Mb (1998)	12	19,099	40%
Sea urchin (*Strongylocentrotus purpuratus*)	814 Mb (2006)	42	~23,500	60%
Thale cress (plant) (*Arabidopsis thaliana*)	140 Mb (2000)	10	~27,500	Not determined
Zebrafish (*Danio rerio*)	1.4 Gb (2013)	50	~41,800	70%
Yeast (*Saccharomyces cerevisiae*)	12 Mb (1996)	32	~5700	30%

Originally adapted from Palladino, M. A. (2006) *Understanding the Human Genome Project*, 2nd ed. Benjamin Cummings.

- **Repetitive sequences.** The presence of introns and the existence of repetitive sequences are two major reasons for the wide range of genome sizes in eukaryotes. In some plants, such as maize, repetitive sequences are the dominant feature of the genome. The maize genome has about 2500 Mb of DNA, more than two-thirds of which is composed of repetitive DNA. In the human, about half of the genome is repetitive DNA.

Comparative Genomics Provides Novel Information about the Genomes of Model Organisms and the Human Genome

As mentioned earlier, in addition to sequencing the human genome, researchers involved in the Human Genome Project also sequenced the genomes of a number of model nonhuman organisms. These include, among others, *E. coli, S. cerevisiae, D. melanogaster*, the nematode roundworm *C. elegans*, and the mouse *M. musculus*. Complete genome sequences of such organisms have been invaluable for comparative genomics studies of gene function in these organisms and in humans. As shown in Table 21.1, the number of genes humans share with other species is very high, ranging from about 30 percent of the genes in yeast to ~80 percent in mice and ~98 percent in chimpanzees. The human genome even contains around 100 genes that are also present in many bacteria.

Comparative genomics has shown us that many genes identified as being involved in human disease are also present in model organisms. For instance, approximately 60 percent of genes associated with nearly 300 human diseases are also present in *Drosophila*. These include genes involved in prostate, colon, and pancreatic cancers; cardiovascular disease; cystic fibrosis; and several other conditions. Here, we consider how comparative genomics studies of several model organisms (sea urchins, dogs, chimpanzees, and Rhesus monkeys) and Neanderthals have revealed interesting elements of the human genome.

The Sea Urchin Genome

In 2006, researchers from the Sea Urchin Genome Sequencing Consortium completed the 814 million bp genome of the sea urchin *Strongylocentrotus purpuratus*. Sea urchins are shallow-water marine invertebrates often studied by developmental biologists. Fossil records indicate that sea urchins appeared during the Early Cambrian period, around 520 million years ago (mya).

A combination of WGS and map-based cloning in BACs was used to complete the genome. Sea urchins have an estimated 23,500 genes, including representative genes for just about all major vertebrate gene families. Sequence alignment and homology searches demonstrate that the sea urchin contains many genes with important functions in humans, yet interestingly, important genes in flies and worms, such as certain cytochrome P450 genes that play a role in the breakdown of toxic compounds, are missing from sea urchins. The sea urchin genome also has an abundance (~25 to 30 percent) of **pseudogenes**—nonfunctional duplications of protein-coding genes. Sea urchins have a smaller average intron size than humans, supporting the general trend revealed by comparative genomics that intron size is correlated with overall genome size.

Urchins have nearly 1000 genes for sensing light and odor, indicative of great sensory abilities. In this respect, their genome is more typical of vertebrates than invertebrates. A number of orthologs of human genes involved in hearing and balance are present, as are many human-disease-associated orthologs, including protein kinases, GTPases, transcription factors, innate immunity, transporters, and low-density lipoprotein receptors. Sea urchins and humans share approximately 7000 orthologs.

The Dog Genome

In 2005 the genome for "man's best friend" was completed, and it revealed that we share about 75 percent of our genes with dogs (*Canis familiaris*), providing a useful model with which to study our own genome. Dogs have a genome that is similar in size to the human genome: about 2.5 billion base pairs with an estimated 18,400 genes.

The dog offers several advantages for studying heritable human diseases. Dogs share many genetic disorders with humans, including over 400 single-gene disorders, sex-chromosome aneuploidies, multifactorial diseases (such as epilepsy), behavioral conditions (such as obsessive-compulsive disorder), and genetic predispositions to cancer, blindness, heart disease, and deafness.

The molecular causes of at least 60 percent of inherited diseases in dogs, such as point mutations and deletions, are similar or identical to those found in humans. In addition, at least 50 percent of the genetic diseases in dogs are breed-specific. Dog breeds resemble isolated human populations in having a small number of founders and a long period of relative genetic isolation. These many similarities make individual dog breeds useful as models of human genetic disorders.

Breeders are now using genetic tests to screen dogs for inherited diseases, for coat color in Labrador retrievers and poodles, and for fur length in Mastiffs. Undoubtedly, we can expect many more genetic tests for dogs in the near future, including DNA analysis for size, type of tail, speed, sense of smell, and other traits deemed important by breeders and owners.

Scientists have been using genomic data to determine the origin of domestic dogs. An analysis of 48,000 markers across the whole genomes of hundreds of dogs and gray wolves from different regions around the world showed that modern dogs shared more sequences in common with Middle Eastern wolves than with Asian wolves. Recent research based on sequence analysis of mitochondrial DNA (mtDNA) from the fossils of ancient dogs and wolves suggests that dogs originated in Europe from gray wolves that are now extinct. In this work, researchers compared mtDNA from samples 1000 to 36,000 years old from 77 dogs of different breeds, 49 wolves, and 4 coyotes. Based on mtDNA sequences, dogs appear to be more closely related to ancient wolves than to modern wolves (a similar conclusion was drawn by a team studying whole genomes). In addition, mtDNA sequences from ancient remains that most closely matched modern dogs were all from European gray wolf samples. Aging of the samples suggests that dog domestication began between 18,800 and 32,100 years ago among hunter-gatherers several thousand years before humans farmed in earnest. Scientists who believe that the domestication of dogs began in Asia have noted that the mtDNA study may be flawed because the researchers were unable to get DNA samples from ancient specimens from the Middle East and from East Asia. Ongoing research in this area will continue, and it will be interesting to see whether genomics can settle the debate about the origins of man's best friend.

The Chimpanzee Genome

The sequence for the chimpanzee (*Pan troglodytes*) genome was completed in 2004. Overall, the chimp and human genome sequences differ by less than 2 percent, and 98 percent of the genes are the same. Comparisons between these genomes offer some interesting insights into what makes some primates humans and others chimpanzees.

The speciation events that separated humans and chimpanzees occurred less than 6.3 mya. Genomic analysis indicates that these species initially diverged but then exchanged genes again before separating completely. Their separate evolution after this point is exhibited in such differences as that seen between the sequence of chimpanzee chromosome 22 and its human ortholog, chromosome 21 (**Table 21.3**; chimps have 48 chromosomes and humans have 46, so their chromosome numbering is different). These chromosomes have accumulated nucleotide substitutions that total 1.44 percent of the sequence. The most surprising difference is the discovery of 68,000 nucleotide insertions or deletions, collectively called *indels***,** in the chimp and human chromosomes, a frequency of 1 indel every 470 bases. In humans, many of these indels are *Alu* insertions in chromosome 21. Although the overall difference in the nucleotide sequence in humans and chimps is small, there are significant differences in what the genes encode. Only 17 percent of the genes analyzed encode identical proteins in both chromosomes; the other 83 percent encode proteins with one or more amino acid differences.

TABLE 21.3 Comparisons between Human Chromosome 21 and Chimpanzee Chromosome 22

	Human 21	Chimpanzee 22
Size (bp)	33,127,944	32,799,845
G + C content	40.94%	41.01%
CpG islands	950	885
SINEs (*Alu* elements)	15,137	15,048
Genes	284	272
Pseudogenes	98	89

Differences in the time and place of gene expression also play a major role in differentiating the two primates. Using DNA microarrays (discussed in Section 21.9), researchers compared expression patterns of 202 genes in human and chimp cells from brain and liver. They found more species-specific differences in expression of brain genes than liver genes. To further examine these differences, Svante Pääbo and colleagues compared expression of 10,000 genes in human and chimpanzee brains and found that 10 percent of genes examined differ in expression in one or more regions of the brain. More importantly, these differences are associated with genes in regions of the human genome that have been duplicated subsequent to the divergence of chimps and humans. This finding indicates that genome evolution, speciation, and gene expression are interconnected. Further work on these segmental duplications and the genes they contain may identify genes that help make us human.

Recently, thanks to sequencing and comparative genomics, researchers have been trying to understand the roles of ~3000 rapidly evolving segments of the human genome known as *human accelerated regions (HARs)*. Smaller than most human genes, HARs average about 250 bp in size. Many HARs function as regulatory sequences such as enhancers. Others encode lncRNAs, and it has been estimated that approximately 5 percent of HARs produce noncoding RNAs.

Many HARs are in close proximity to genes that control developmental processes, and scientists hypothesize they may have roles in human development. For example, certain HARs and the genes they regulate have been found in regions of the human brain that are larger and more well developed than in chimpanzees. Experiments involving overexpression of these HARs results in mice with developmental characteristics of specific regions of the brain resembling human brain development.

The Rhesus Monkey Genome

The Rhesus macaque monkey (*Macaca mulatta*), another primate, has served as one of the most important model organisms in biomedical research. Macaques have played central roles in our understanding of cardiovascular disease, aging, diabetes, cancer, depression, osteoporosis, and many other aspects of human health. They have been essential for research on AIDS vaccines and for the development of polio vaccines.

The macaque's genome is the first monkey genome to have been sequenced. A main reason geneticists are so excited about the completion of this sequencing project is that macaques provide a more distant evolutionary window that is ideally suited for comparing and analyzing human and chimpanzee genomes. Whereas humans and chimpanzees shared a common ancestor approximately 6.3 mya, macaques split from the ape lineage that led to chimpanzees and humans about 25 mya. The macaque and human genome have thus diverged farther from one another, as evidenced by the ~93 percent sequence identity between humans and macaques compared to the ~98 percent sequence identity shared by humans and chimpanzees.

The macaque genome was published in 2007, and it was no surprise to learn that it consists of 2.87 billion bp (similar in size to the human genome) contained in 22 chromosomes (20 autosomes, an X, and a Y) with ~20,000 protein-coding genes. While comparative analyses of this genome are ongoing, a number of interesting features have already been revealed. As in humans, about 50 percent of the genome consists of repeat elements (transposons, LINEs, SINEs). Gene duplications and gene families are abundant, including cancer gene families found in humans.

A number of interesting surprises have also been uncovered. For instance, recall from earlier in the text (see Chapter 14) our discussion about the genetic disorder phenylketonuria (PKU), an autosomal recessive inherited condition where individuals cannot metabolize the amino acid phenylalanine due to mutation of the phenylalanine hydroxylase (*PAH*) gene. The histidine substitution encoded by a mutation in the *PAH* gene of humans with PKU appears as the wild-type amino acid in the protein from macaques with the same gene mutation. But why macaques with this mutation do not develop PKU is not known. Further analysis of the macaque genome and comparison to the human and chimpanzee genome will be invaluable for geneticists studying genetic variations that played a role in primate evolution.

The Neanderthal Genome and Modern Humans

In 2010, a team of scientists led by Svante Pääbo at the Max Planck Institute for Evolutionary Anthropology in Germany and the U.S. biotechnology company 454 Life Sciences reported completion of a rough draft of the Neanderthal (*Homo neanderthalensis*) genome encompassing more than 3 billion bp of Neanderthal DNA and about two-thirds of the genome. Previously, in 1997, Pääbo's lab sequenced portions of Neanderthal mitochondrial DNA from a fossil. In late 2006, Pääbo's group along with a number of scientists in the United States reported the first sequence of ~65,000 bp of nuclear DNA isolated from Neanderthal bone samples. Bones from three females who lived in Vindija Cave in Croatia about 38,000–44,000 years ago were used to produce the draft sequence of the Neanderthal nuclear genome. A sequence recently recovered from calcified remains of a Neanderthal that lived over 200,000 years ago is thought to be the oldest Neanderthal DNA ever analyzed.

Because Neanderthals are members of the human family, and closer relatives to humans than chimpanzees, the Neanderthal genome provides an unprecedented opportunity to use comparative genomics to advance our understanding of evolutionary relationships between modern humans and our predecessors. In particular, scientists are interested in identifying areas in the genome where humans have undergone rapid evolution since splitting (diverging) from Neanderthals. Much of this analysis involves a comparative genomics approach to compare the Neanderthal genome to the human and chimpanzee genomes.

The human and Neanderthal genomes are 99 percent identical. Comparative genomics has identified 78 protein-coding sequences in humans that seem to have arisen since the divergence from Neanderthals and that may have helped modern humans adapt. Some of these sequences are involved in cognitive development and sperm motility. Of the many genes shared by humans and Neanderthals, *FOXP2* is a gene that has been linked to speech and language ability. There are many genes that influence speech, so this finding does not mean that Neanderthals spoke as we do. But because Neanderthals had the same modern human *FOXP2* gene scientists have speculated that Neanderthals possessed linguistic abilities.

The realization that modern humans and Neanderthals lived in overlapping ranges as recently as 30,000 years ago has led to speculation about the interactions between modern humans and Neanderthals. Genome studies suggest that interbreeding took place between Neanderthals and modern humans an estimated 45,000–80,000 years ago in the eastern Mediterranean. In fact, the genome of non-African

H. sapiens contains approximately 1–4 percent of a sequence inherited from Neanderthals. These exciting studies, previously thought to be impossible, are having ramifications in many areas of the study of human evolution, and it will be interesting indeed to follow the progress of this work.

21.8 Metagenomics Applies Genomics Techniques to Environmental Samples

Metagenomics, also called **environmental genomics,** is the discipline that uses whole-genome shotgun approaches to sequence genomes from entire communities of microbes in natural environments, which includes living organisms and environmental samples of water, air, and soil. Oceans, glaciers, deserts, and virtually every other environment on Earth are being sampled for metagenomics projects. Human genome pioneer J. Craig Venter left Celera to form the J. Craig Venter Institute, and his group played a central role in developing metagenomics as an emerging area of genomics research.

One of the institute's major initiatives was a global expedition to sample marine and terrestrial microorganisms from around the world and to sequence their genomes, called the *Sorcerer II* Global Ocean Sampling (GOS) Expedition. Between 2004 and 2006, Venter and his researchers traveled the globe by yacht, in a sailing voyage covering 32,000 nautical miles and described as a modern-day version of Charles Darwin's famous voyage on the *H.M.S. Beagle*. One of the earliest expeditions by this group sequenced bacterial genomes from the Sargasso Sea off Bermuda. This project yielded over 1.2 million novel DNA sequences from 1800 microbial species, including 148 previously unknown bacterial species, and identified hundreds of photoreceptor genes.

A key benefit of metagenomics is its potential for teaching us more about millions of yet uncharacterized species of bacteria. Many new viruses, particularly bacteriophages, have been identified through metagenomics studies of water and soil samples. Metagenomics is providing important new information about genetic diversity in microbes that is key to understanding complex interactions between microbial communities and their environment, as well as allowing phylogenetic classification of newly identified microbes. Metagenomics also has great potential for identifying genes with novel functions, some of which may have valuable applications in medicine and biotechnology.

The general method used in metagenomics often involves isolating DNA directly from an environmental sample without requiring cultures of the microbes or viruses. Such an approach is necessary because often it is difficult to replicate the complex array of growth conditions the microbes need to survive in culture.

An example of how metagenomics can provide novel insight into the microbial world around us is reflected by a recent study of the microbiota found in New York City subways. This project revealed that most microbes present are non-disease-causing bacteria normally prevalent on human skin and in the GI tract. Although, occasionally, pathogens such as *Bacillus anthracis* were identified. But almost half of the DNA sequenced did not match any organism in eukaryotic, bacterial, archaeal and viral genome databases!

The Human Microbiome Project

In 2007 the National Institutes of Health announced plans for the **Human Microbiome Project (HMP),** a $170 million effort to sequence the genomes of an estimated 600–1000 bacteria, viruses, yeasts, and other microorganisms that live on and inside humans. At the start of the project, microorganisms were thought to outnumber human cells by about 10 to 1, although this is likely a prediction that is too high.

Many microbes, such as *E. coli* in the digestive tract, have important roles in human health, and of course other microbes make us ill. The HMP had several major goals, including:

- Determining if individuals share a core human microbiome.
- Understanding whether changes in the microbiome can be correlated with changes in human health.
- Developing new methods, including bioinformatics tools, to support analysis of the microbiome.
- Addressing ethical, legal, and social implications raised by human microbiome research.

The HMP involved about 200 scientists at 80 institutions. In 2012 a series of papers were published summarizing findings from the HMP. The HMP analyzed 15 body sites from males and 18 sites from females from 242 healthy individuals in the United States and applied WGS of genomes for the microbes and viruses present at these sites. Each person was sampled up to three times over nearly two years. In addition to WGS analysis, sequences for 16*S* rRNA genes were used specifically to compare bacterial samples. More than 3000 reference sequences for microbes isolated from the human body were developed.

The HMP amassed more than 1000 times the sequencing data generated by the HGP. We have formulated the following concepts about the human microbiome:

- Sequence data from the HMP have identified an estimated 81 to 99 percent of the microbes and viruses distributed among body areas in human males and females.
- As many as 1000 bacterial strains may be present in each person.

- The microbiome starts at birth. Babies acquire bacteria from their mothers' microbiome.
- A surprise to HMP scientists, the microbiome can be substantially different from person to person. Based on variation between individuals, an estimated 10,000 bacterial species may be part of the total human microbiome.
- Although the microbiome of the human gut differs from person to person, it remains relatively stable over time in individuals.

Based on these findings, there is no single reference human microbiome to which people can be compared. Microbial diversity varies greatly from individual to individual, and a personalization of the microbiome occurs in individuals. For instance, comparing sequences of the microbiomes from two healthy people of equivalent age reveals microbiomes that can be quite different. There are, however, similarities in certain parts of the body, with signature bacteria and characteristic genes associated that are specific to certain locations.

Knowledge about the personalized nature of the microbiome is already proving valuable for improving human health and medicine, which in the future may include microbiome-specific therapeutic drugs. Criteria are being sought for a healthy microbiome, which is expected to help determine the role of bacteria in maintaining normal health. This may provide insight into how antibiotics can disturb a person's microbiome and why certain individuals are susceptible to certain diseases, especially chronic conditions such as psoriasis, irritable bowel syndrome, and potentially even obesity.

Related to this project, a team of researchers at the University of California, Los Angeles, analyzed DNA sequences from 101 college students, 49 of whom had acne and 52 of whom did not. Over 1000 strains of *Propionibacterium acnes* were isolated. Using WGS and bioinformatics, researchers clustered these strains into ten strain types (related strains). Six of these types were more common among acne-prone students, and one type appeared repeatedly in skin samples from students without acne. Sequence analysis of types associated with acne indicated groups of genes that may contribute to the skin disease. Further analysis of these strain types may help dermatologists develop new drugs targeted at killing acne-causing strains of *P. acnes*.

A *Venn diagram*, like the image shown in **Figure 21.13**, is a common way to represent overlapping data in metagenomics datasets. In this figure, overlapping circles indicate numbers of human gut microbiome genes from individuals with liver cirrhosis, Type 2 diabetes, and irritable bowel syndrome. Notice that each disease has a unique profile of microbial genes but that significant overlaps between microbial genes for each disease occur. Of the microbial genes, 403 were shared and thus considered common markers for all three diseases.

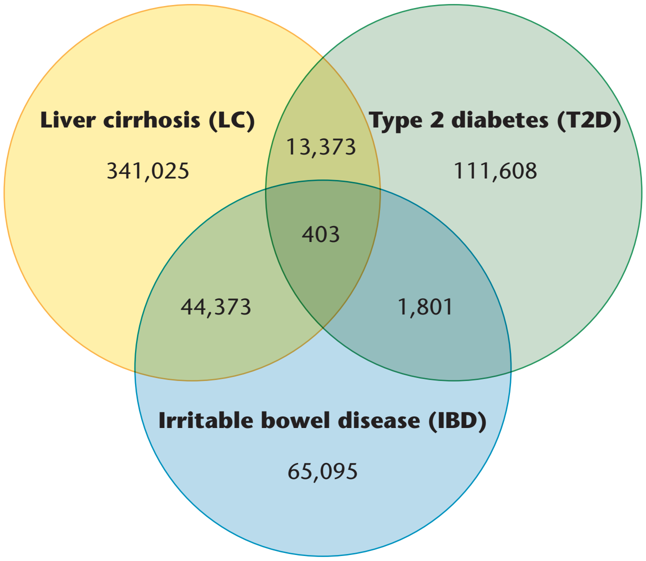


FIGURE 21.13 Venn diagram representation of gut microbial genes from patients with liver cirrhosis, Type 2 diabetes, and irritable bowel syndrome. Notice that different diseases show large numbers of unique genes with smaller numbers of shared genes.

The Case Study at the end of this chapter briefly discusses a clinical application focused on the importance of gut microbes for intestinal health.

21.9 Transcriptome Analysis Reveals Profiles of Expressed Genes in Cells and Tissues

Once any genome has been sequenced and annotated, a formidable challenge remains: that of understanding genome function by analyzing the genes it contains and the ways the genes expressed in the genome are regulated. **Transcriptome analysis** (also called **transcriptomics** or global analysis of gene expression) studies the expression of genes in a genome both qualitatively—by identifying which genes are expressed or not expressed—and quantitatively—by measuring varying levels of expression for different genes. In other words, transcriptome analysis attempts to catalog and quantify the total RNA content of a cell, tissue, or organism.

Even though in theory all cells or tissue types of an organism possess the same genes, depending on location certain genes will be highly expressed, others expressed at low levels, and some not expressed at all. Transcriptome analysis reveals gene-expression profiles that, for the same genome, may vary from cell to cell or from tissue type to tissue type. Identifying where and when genes are expressed by a genome is essential for understanding how the genome functions.

Transcriptome analysis provides insights into (1) normal patterns of gene expression that are important for understanding how a cell or tissue type differentiates during development, (2) how gene expression dictates and controls the physiology of differentiated cells, and (3) mechanisms of disease development that result from or cause gene-expression changes in cells. Later in the text (see Chapter 22), we will consider why transcriptome analysis is gradually becoming an important diagnostic tool in certain areas of medicine. For example, examining gene-expression profiles in a cancerous tumor can help diagnose tumor type, determine the likelihood of tumor metastasis (spreading), and develop the most effective treatment strategy.

DNA Microarray Analysis

A number of different techniques can be used for transcriptome analysis. PCR-based methods—such as reverse transcription PCR (RT-PCR) and quantitative real-time PCR (qPCR) (described in Chapter 20) are useful because of their ability to detect genes expressed at low levels. For nearly two decades **DNA microarray analysis** has been widely used because it enables researchers to analyze all of a sample's expressed genes simultaneously (see Figure 21.14).

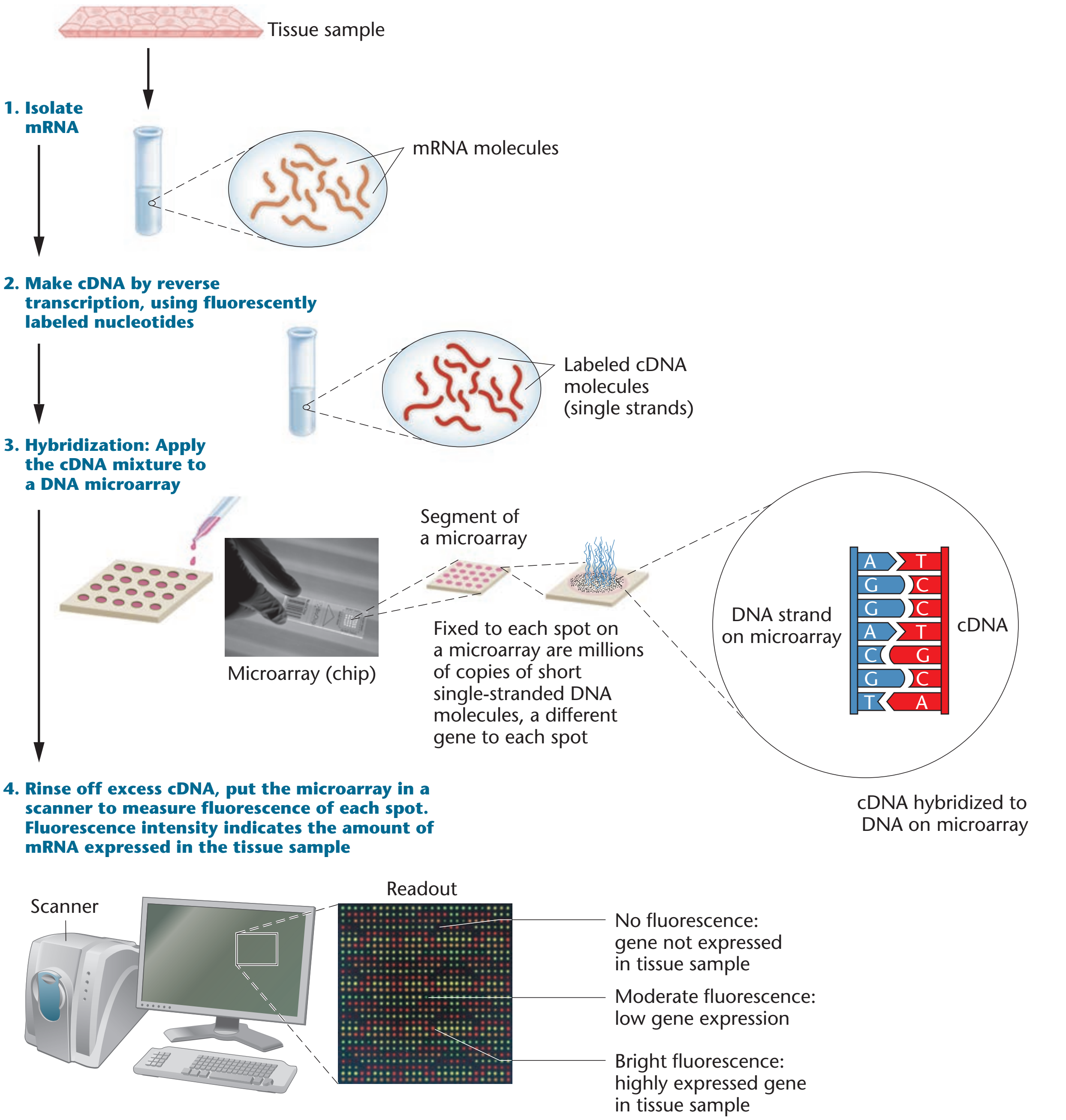

Tissue sample
1. Isolate mRNA
mRNA molecules
2. Make cDNA by reverse transcription, using fluorescently labeled nucleotides
Labeled cDNA molecules (single strands)
3. Hybridization: Apply the cDNA mixture to a DNA microarray
Segment of a microarray
Microarray (chip)
Fixed to each spot on a microarray are millions of copies of short single-stranded DNA molecules, a different gene to each spot
DNA strand on microarray
A G G A C G T
T C C T G C A
cDNA
cDNA hybridized to DNA on microarray
4. Rinse off excess cDNA, put the microarray in a scanner to measure fluorescence of each spot. Fluorescence intensity indicates the amount of mRNA expressed in the tissue sample
Scanner
Readout
No fluorescence: gene not expressed in tissue sample
Moderate fluorescence: low gene expression
Bright fluorescence: highly expressed gene in tissue sample

FIGURE 21.14 DNA microarray analysis for analyzing gene-expression patterns in a tissue.

Most DNA microarrays, also known as **gene chips,** are prepared by "spotting" single-stranded DNA molecules onto glass slides using a computer-controlled high-speed robotic arm called an arrayer. Arrayers are fitted with a number of tiny pins. Each pin is immersed in a small amount of solution containing millions of copies of a different single-stranded DNA molecule. [For example, many microarrays are prepared with single-stranded complementary DNA (cDNAs) or expressed sequence tags (ESTs)—short fragments of DNA cloned from expressed genes.] The arrayer fixes the DNA onto the slide at specific locations (called spots, fields, or features) that will be scanned and recorded by a computer. A single microarray can have over 20,000 different spots of DNA (and over 1 million for exon-specific microarrays), each containing a unique sequence that serves as a probe for a different gene.

To use a microarray for transcriptome analysis, scientists typically begin by extracting mRNA from cells or tissues (Figure 21.14). The mRNA is usually then reverse transcribed to synthesize cDNA tagged with fluorescently labeled nucleotides. Microarray studies often involve comparing gene expression in different cell or tissue samples. cDNA prepared from one tissue is usually labeled with one color dye, red for example, and cDNA from another tissue is labeled with a different-colored dye, such as green. Labeled cDNAs are then denatured and incubated overnight with the microarray so that they will hybridize to spots on the microarray that contain complementary DNA sequences. Next, the microarray is washed, and then it is scanned by a laser that causes the cDNA hybridized to the microarray to fluoresce. The patterns of fluorescent spots reveal which genes are expressed in the tissue of interest, and the intensity of spot fluorescence indicates the relative level of expression. The brighter the spot, the more the particular mRNA is expressed in that tissue.

DNA microarrays have dramatically changed the way gene-expression patterns are analyzed. As discussed earlier (in Chapter 20), Northern blot analysis was one of the earliest methods used for analyzing gene expression. Then PCR techniques proved to be more rapid and have increased sensitivity. The biggest advantage of DNA microarrays is that they enable thousands of genes to be studied simultaneously. As a result, however, they can generate an overwhelming amount of gene-expression data. Over 1 million human gene-expression datasets are available in publicly accessible databases, and commercially available DNA microarrays for analyzing human gene expression are now widely used (see **Figure 21.15**).

But one limitation of DNA microarrays is that they can often yield variable results. For example, one experiment under certain conditions may not always yield similar patterns of gene expression when the experiment is repeated. Some of this variability can be due to real differences in gene expression, but others can be the result of variation in chip preparation, cDNA synthesis, probe hybridization, or washing conditions, all of which must be carefully controlled to limit such variability. Commercially available DNA microarrays can reduce the variability that can result when individual researchers make their own arrays.

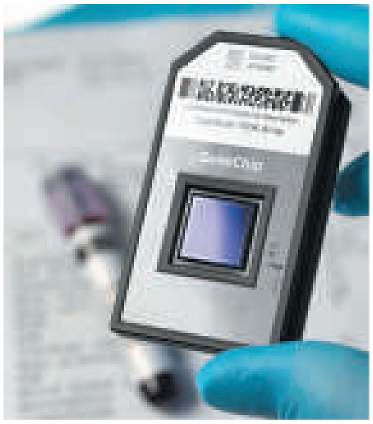


FIGURE 21.15 A commercially available DNA microarray, called a GeneChip®, marketed by Affymetrix, Inc. This particular microarray can be used to analyze expression for approximately 50,000 RNA transcripts. It contains 22 different probes for each transcript and allows scientists to simultaneously assess the expression levels of most of the genes in the human genome.

As you will learn later (in Chapter 22), researchers are also using DNA microarrays to compare patterns of gene expression in tissues in response to different conditions, to compare gene-expression patterns in normal and diseased tissues, and to identify pathogens.

As we have discussed, the significant value of gene expression microarrays has been the ability to quantify RNA expression for large numbers of genes simultaneously.

Hybridization-based microarrays provided the first relatively inexpensive way to detect and quantify transcripts on a large scale. However, the next section describes an even more modern approach that will likely render DNA microarrays obsolete in the future.

RNA Sequencing Technology Allows for *In Situ* Analysis of Gene Expression

In recent years, significant progress has been made on direct **RNA sequencing (RNA-seq),** also called whole-transcriptome shotgun sequencing. One limitation of microarrays is that the investigator is limited to study the expression of only those genes with probes on the chip. Conversely, RNA-seq not only allows for quantitative analysis of all RNAs expressed in a particular tissue, but it also provides actual sequence data. And RNA-seq can also be carried out inside the cell (*in situ*). For this application of RNA-seq, the cell itself can serve as a "gene chip."

Generally, most RNA-seq methods incorporate reverse-transcribing RNA *in situ*, sequencing cDNA, mapping sequences to a reference genome (to determine which sequences were transcribed from specific genes), and quantifying gene expression. Methods incorporating fluorescence *in situ* RNA-seq are enabling scientists to visualize where specific RNAs are being transcribed in intact cells and tissues.

It is now also becoming possible to carry out DNA sequencing and RNA-seq on *individual* cells! Several groups are taking an integrated approach to sequence genomes and transcriptomes in the same cell. This strategy enables researchers to correlate genetic variability and mRNA expression variability simultaneous in single cells—thus analyzing both the genome and the transcriptome.

We will consider applications of RNA-seq later in the text (see Chapter 22), but clearly this approach has already demonstrated great value for disease diagnosis including understanding how genome variations such as CNVs impact transcriptome expression. Now that we have discussed genomes and transcriptomes, we turn our attention to the ultimate end products of most genes, the proteins encoded by a genome.

21.10 Proteomics Identifies and Analyzes the Protein Composition of Cells

As genomes have been sequenced and studied, biologists have focused increasingly on understanding the complex structures, functions, and interactions of the proteins that genomes encode—the **proteome,** defined as the complete set of proteins encoded by a given genome. **Proteomics** is the identification, characterization, and quantitative analysis of proteomes.

Proteomics provides information about many things:

- A protein's structure and function
- Posttranslational modifications
- Protein—protein, protein—nucleic acid, and protein—metabolite interactions
- Cellular localization of proteins
- Protein stability and aspects of translational and posttranslational levels of gene-expression regulation
- Relationships (shared domains, evolutionary history) to other proteins

Proteomics projects have been used to characterize major families of proteins for some species. For example, about two-thirds of the *Drosophila* proteome has been well cataloged using proteomics.

Proteomics is also of clinical value because it allows comparison of proteins in normal and diseased tissues, which can lead to the identification of proteins as biomarkers for disease conditions. Proteomic analysis of mitochondrial proteins during aging, proteomic maps of atherosclerotic plaques from human coronary arteries, and protein profiles in saliva as a way to detect and diagnose diseases are examples of such work.

Reconciling the Number of Genes and the Number of Proteins Expressed by a Cell or Tissue

While Beadle and Tatum's one gene: one enzyme hypothesis was a worthy proposal in the 1940s (see Chapter 14), genomics has revealed that the link between gene and gene product is often much more complex. Genes can have multiple transcription start sites that produce several different types of RNA transcripts. Alternative splicing and editing of pre-mRNA molecules can generate dozens of different proteins from a single gene. As a result, proteomes are substantially larger than genomes. Sequencing of mRNAs from human tissues found that over 95 percent of protein-coding genes with more than one exon are alternatively spliced.

However, it is unclear how many different proteins are translated from this pool of transcripts. To address this, the Human Proteome Map (HPM), published in 2014, aimed to catalog the human proteome in all its complexity. This project involved proteomic analysis of a wide range of human tissues and cell types using methods we will discuss in the next section. Based on the results of this study, we now know that the ~20,000 protein-coding genes in the human genome can produce at least 290,000 different proteins. The HPM accounted for ~85% of all annotated protein-coding genes in humans that currently exist in human proteomics databases. Refer to PDQ 20 to access the online database for the HPM.

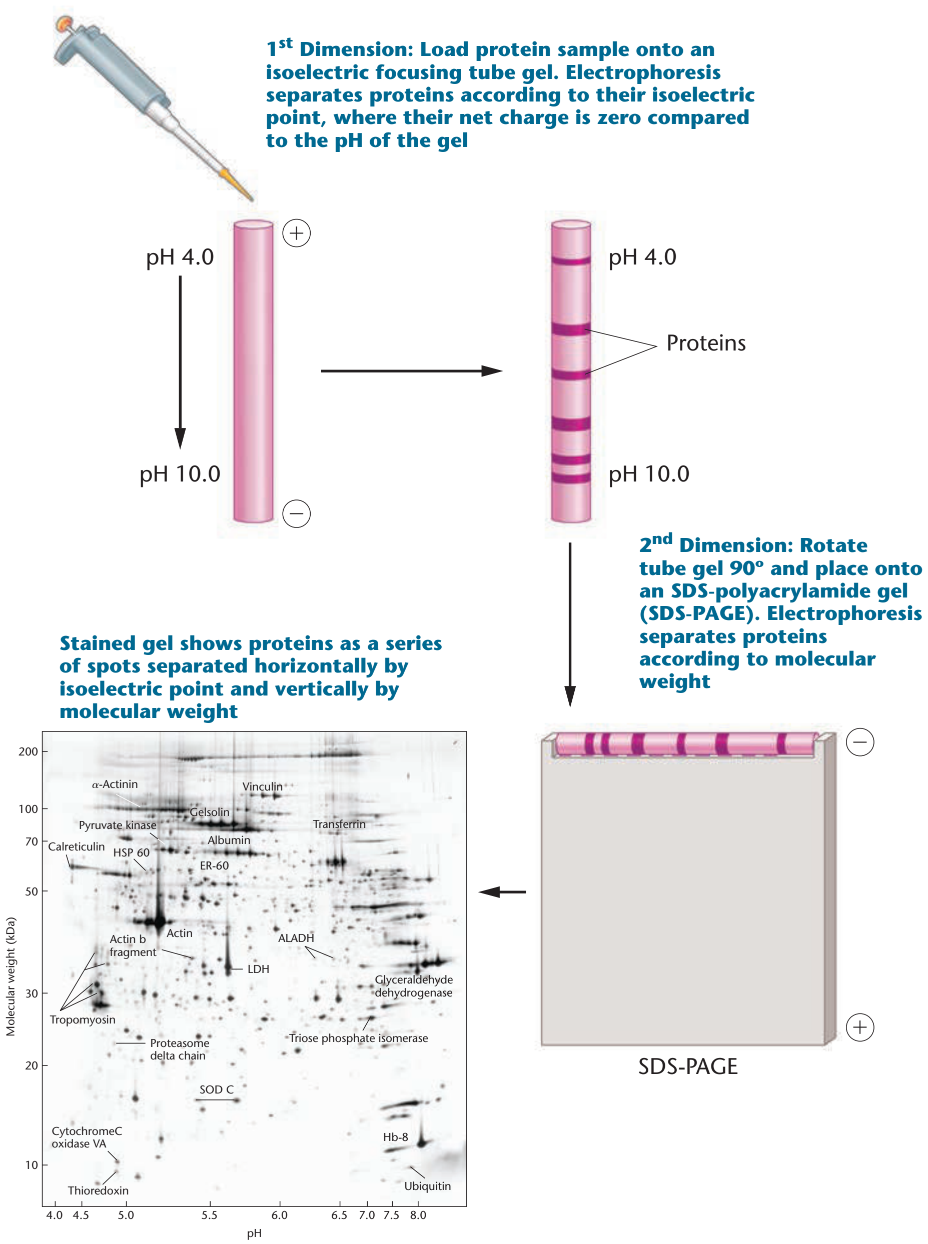
The specific protein content (or profile) of a cell is determined in large part by its gene-expression patterns—its transcriptome. However, a number of other factors affect the proteome profile of a cell. To begin with, many proteins undergo co-translational or posttranslational modifications, such as cleavage of a signal sequence that targets a protein for an organelle pathway, a propeptide, or initiator methionine residues; linkage to carbohydrates and lipids; or the addition of chemical groups through methylation, acetylation, and phosphorylation; and other modifications. Over a hundred different mechanisms of posttranslational modification are known.

In addition, many proteins work via elaborate protein—protein interactions or as part of a large macromolecular complex. Furthermore, although every cell in the body contains an equivalent set of genes, not all cells express the same genes and, hence, the same proteins. Proteomics also considers proteins that a cell might acquire from another cell, not just the proteins encoded by the genome of the cell type being analyzed.

The early history of proteomics dates back to 1975 and the development of **two-dimensional gel electrophoresis (2DGE),** a technique for separating hundreds to thousands of proteins with high resolution. In this technique, proteins isolated from cells or tissues of interest are first loaded onto a polyacrylamide tube gel and separated by *isoelectric focusing*, which causes proteins to migrate based on their electrical charge in a pH gradient. During isoelectric focusing, proteins migrate until they reach the location in the gel where their net charge is zero relative to the pH of the gel. Then in a second migration, perpendicular to the first, the proteins are separated by their molecular weight using *sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)* (**Figure 21.16**).

It is not uncommon for a 2D gel loaded with a complex mixture of proteins to show several thousand spots, as in Figure 21.16, which displays the complex mixture of proteins in human platelets (thrombocytes). Particularly abundant protein spots in this gel have been labeled with the names of identified proteins. With thousands of different spots on the gel, how are the identities of the proteins ascertained?

In some cases, 2D gel patterns from experimental samples can be compared to gels run with reference standards containing known proteins with well-characterized migration patterns. Reference gels for different biological samples such as human plasma are available, and computer software



1st Dimension: Load protein sample onto an isoelectric focusing tube gel. Electrophoresis separates proteins according to their isoelectric point, where their net charge is zero compared to the pH of the gel

2nd Dimension: Rotate tube gel 90° and place onto an SDS-polyacrylamide gel (SDS-PAGE). Electrophoresis separates proteins according to molecular weight

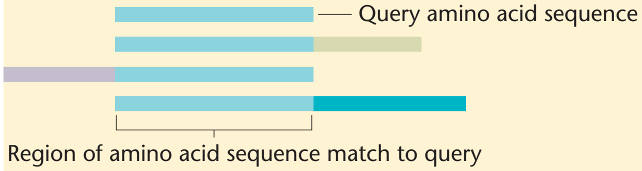
Stained gel shows proteins as a series of spots separated horizontally by isoelectric point and vertically by molecular weight

FIGURE 21.16 Two-dimensional gel electrophoresis (2DGE) is a useful method for separating proteins in a protein extract from cells or tissues that contains a complex mixture of proteins with different biochemical properties. The two-dimensional gel photo shows separations of human platelet proteins. Each spot represents a different polypeptide separated by isoelectric point, pH (*x*-axis), and molecular weight (*y*-axis). In this photo, some protein spots have been identified by name based on comparison to a reference gel or by determination of a protein sequence using mass spectrometry. Notice that many spots on the gel are unlabeled, indicating proteins of unknown identity.

programs can be used to align and compare the spots from different gels. In the early days of 2DGE, proteins were often identified by cutting spots out of a gel and sequencing the amino acids the spots contained. Only relatively small sequences of amino acids can typically be generated this way; rarely can an entire polypeptide be sequenced using this technique. BLAST and similar programs can be used to search protein databases containing amino acid sequences of known proteins. However, because of alternative splicing or posttranslational modifications, peptide sequences may not always match easily with the final product, and the identity of the protein may have to be confirmed by another approach.

NOW SOLVE THIS

21.2 Annotation of a proteome attempts to relate each protein to a function in time and space. Traditionally, protein annotation depended on an amino acid sequence comparison between a query protein and a protein with known function. If the two proteins shared a considerable portion of their sequence, the query would be assumed to share the function of the annotated protein. Following is a representation of this method of protein annotation involving a query sequence and three different human proteins. Note that the query sequence aligns to common domains within the three other proteins. What argument might you present to suggest that the function of the query is not related to the function of the other three proteins?


Query amino acid sequence
Region of amino acid sequence match to query

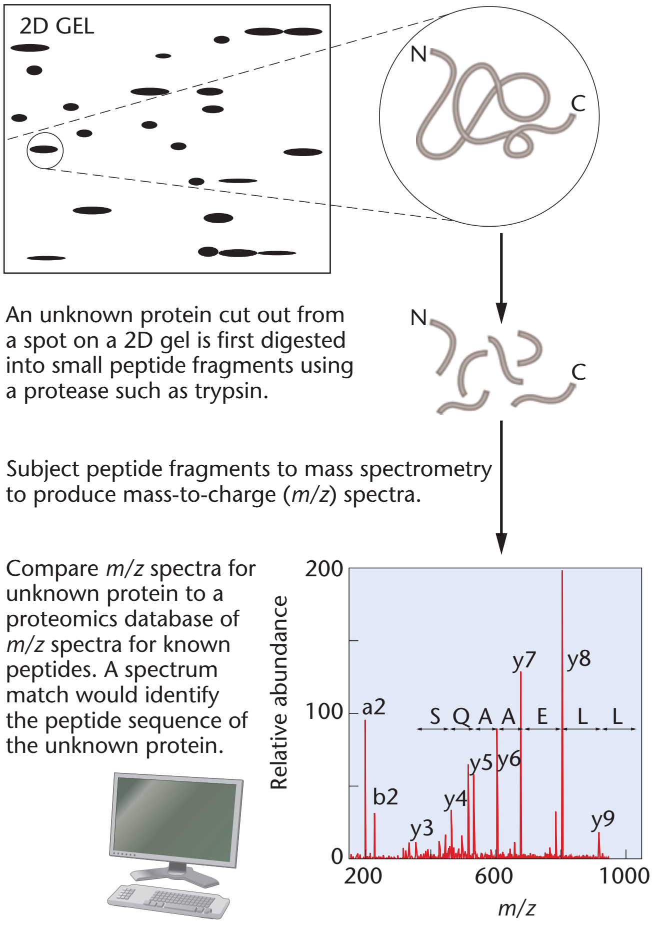
■ **HINT:** *This problem asks you to think about sequence similarities between four proteins and predict functional relationships. The key to its solution is to remember that although protein domains may have related functions, proteins can contain several different interacting domains that determine protein function.*

For more practice, see Problems 22 and 24.

Mass Spectrometry for Protein Identification

As important as 2DGE has been for protein analysis, **mass spectrometry (MS)** has been instrumental to the development of proteomics. Mass spectrometry techniques analyze ionized samples in gaseous form and measure the *mass-to-charge (m/z) ratio* of the different ions in a sample. Proteins analyzed by mass spectrometry generate *m/z* spectra that can be correlated with an *m/z* database containing known protein sequences to discover the protein's identity. Certain forms of MS can provide peptide sequences directly from spectra. Some of the most valuable proteomics applications of this technology are to identify an unknown protein or proteins in a complex mix of proteins, to sequence peptides, to identify posttranslational modifications of proteins, and to characterize multiprotein complexes. Many other biochemical methods can be used together with MS, and new MS techniques do not involve running gels. Here we simply provide an introduction to MS.

One common MS approach is *matrix-assisted laser desorption ionization (MALDI)*. This approach is ideally suited for identifying proteins and is widely used for proteomic analysis of tissue samples. In brief, MALDI employs an ultraviolet laser to heat, vaporize, and ionize peptide fragments. Released ions are then analyzed for mass; MALDI displays the *m/z* ratio of each ionized peptide as a series of peaks representative of the molecular masses of peptides in the mixture and their relative abundance (**Figure 21.17**). Because different proteins produce different sets of peptide fragments, MALDI produces a peptide "fingerprint" that is characteristic of the protein being analyzed.



An unknown protein cut out from a spot on a 2D gel is first digested into small peptide fragments using a protease such as trypsin.

Subject peptide fragments to mass spectrometry to produce mass-to-charge (*m/z*) spectra.

Compare *m/z* spectra for unknown protein to a proteomics database of *m/z* spectra for known peptides. A spectrum match would identify the peptide sequence of the unknown protein.

FIGURE 21.17 Mass spectrometry for identifying an unknown protein isolated from a 2D gel. The peptide in this example was revealed to have the amino acid sequence serine (S)-glutamine (Q)-alanine (A)-alanine (A)-glutamic acid (E)-leucine (L)-leucine (L), shown in single-letter amino acid code.

For many MS approaches including MALDI, proteins are first extracted from cells or tissues of interest and usually separated by 2DGE. Protein spots are cut out of the gel, and proteins are purified out of each gel spot. Computer-automated high-throughput instruments are available that can pick all the spots out of a 2D gel, after which MALDI is used to identify the proteins in the different spots. Just about any source providing a sufficient number of cells can be used: blood, whole tissues and organs, tumor samples, microbes, and many other substances. Many proteins involved in cancer have been identified by the use of MALDI to compare protein profiles in normal tissue and tumor samples.

Databases of MALDI-generated *m/z* spectra for different peptides can be analyzed to look for matches between *m/z* spectra of unknown samples and those of known proteins. One limitation of this approach is database quality. An unknown protein from a 2D gel can only be identified by MALDI if proteomics databases have a MALDI spectrum for that protein. But as is occurring with genomics databases, proteomics databases with thousands of well-characterized proteins from different organisms are rapidly developing. The human proteome database described earlier in this section was developed from MS data.

As mentioned earlier, DNA has been recovered from fossils, but the general assumption has been that proteins degrade in fossilized materials and cannot be recovered. Mass spectrometry analysis of bone tissue from a *Tyrannosaurus rex* skeleton excavated from the Hell Creek Formation in eastern Montana and estimated to be 68 million years old is a good demonstration that fossilization does not fully destroy all proteins in well-preserved fossils under certain conditions. This research also demonstrates the power and sensitivity of mass spectrometry as a proteomics tool.

In this work, medullary tissue was removed from the femur bones. Medullary tissue is porous, spongy bone that contains bone marrow cells, blood vessels, and nerves. *T. rex* proteins extracted from the bone tissue showed cross-reactivity with antibodies to chicken collagen and were digested by the collagen-specific protease collagenase. These results suggested that the *T. rex* protein samples contained collagen, a major matrix component of bone, ligaments, tendons, and skin.

To definitively identify the presence of collagen, tryptic peptides from the *T. rex* samples were analyzed by mass spectrometry. The *m/z* spectra for one of the *T. rex* peptides was identified from a database of *m/z* spectra as corresponding to collagen. Compare the spectrum for the *T. rex* peptide in **Figure 21.18(a)** to that of a synthetic

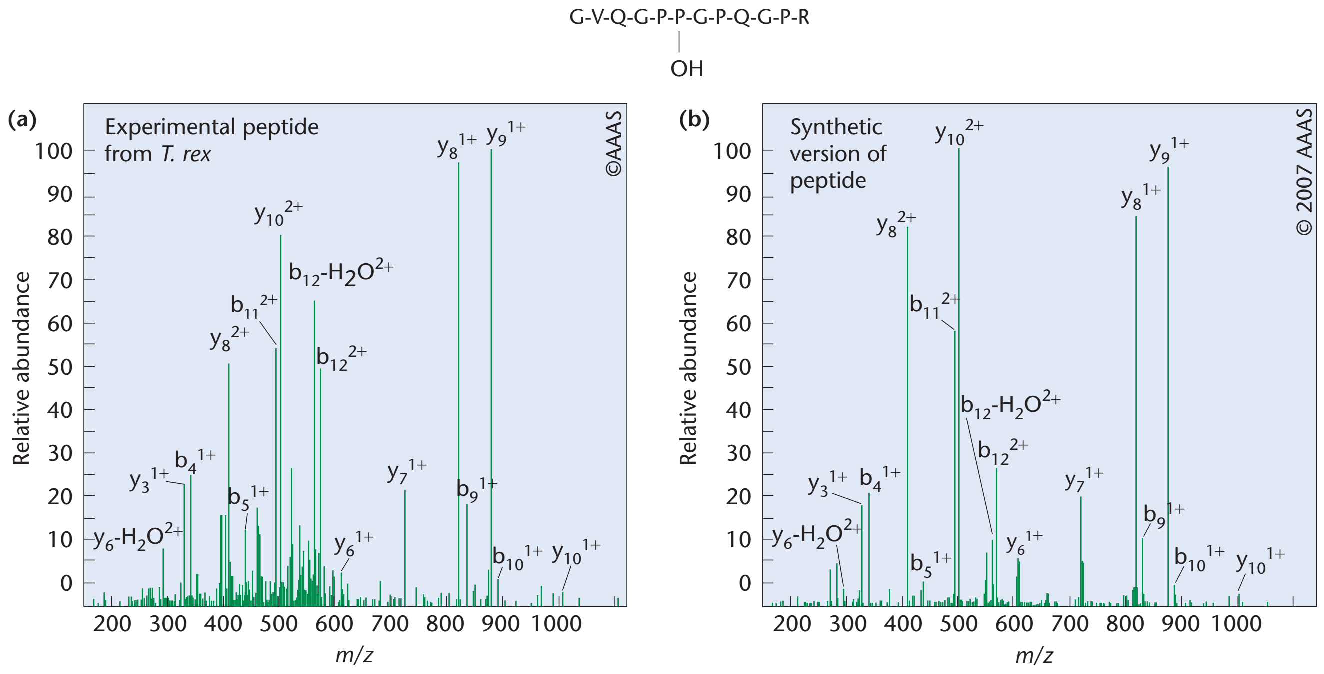


FIGURE 21.18 Mass spectrometry (MS) fragmentation patterns for a peptide sequence—GVQGPP(OH)GPQGPR—from *T. rex*. The peptide sequence, contains a charged hydroxyl group characteristic of collagen. Mass spectrometry of a synthetic version of collagen peptide shows good alignment with the *m/z* spectra for fragmented ions from the *T. rex* peptide, thus confirming the *T. rex* sequence as collagen and demonstrating the value of MS techniques.

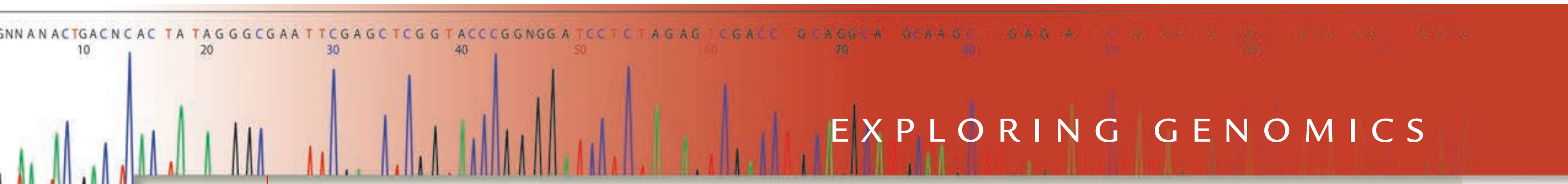
version of a collagen peptide created for sequence validation [**Figure 21.18(b)**]. You will notice that the *m/z* ratios for all major fragmentation ions are in very good alignment with those from the synthetic version, confirming that the *T. rex* sequence is collagen. The *T. rex* peptide also contained a hydroxyl group attached to a proline residue. Proline hydroxylation is a characteristic feature of collagen. In addition, the amino acid sequence of the *T. rex* peptide aligned with an isoform of chicken collagen, further demonstrating sequence similarity. Such work has provided excellent experimental evidence to support the widely accepted theory that birds and dinosaurs are close relatives.

Similar results were obtained for peptides from a 160,000- to 600,000-year-old mastodon (*Mammut americanum*) that showed matches to collagen from extant species, including collagen isoforms from humans, chimps, dogs, cows, chickens, elephants, and mice.

NOW SOLVE THIS

21.3 Because of its accessibility and biological significance, the proteome of human plasma has been intensively studied and used to provide biomarkers for such conditions as myocardial infarction (troponin) and congestive heart failure (B-type natriuretic peptide). Polanski and Anderson compiled a list of 1261 proteins, some occurring in plasma, that appear to be differentially expressed in human cancers [Polanski, M., and Anderson, N. L. (2006) *Biomarker Insights* 2:1–48]. Of these 1261 proteins, only 9 have been recognized by the FDA as tumor-associated proteins. First, what advantage should there be in using plasma as a diagnostic screen for cancer? Second, what criteria should be used to validate that a cancerous state can be assessed through the plasma proteome?

- **HINT:** *This problem asks you to consider criteria that are valuable for using plasma proteomics as a diagnostic screen for cancer. The key to its solution is to consider proteomics data that you would want to evaluate to determine whether a particular protein is involved in cancer.*



EXPLORING GENOMICS

Contigs, Shotgun Sequencing, and Comparative Genomics

Mastering Genetics Visit the Study Area: Exploring Genomics

In this chapter, we discussed how WGS can be used to assemble chromosome maps. Recall that in this technique chromosomal DNA is digested with different restriction enzymes (or mechanically sheared) to create a series of overlapping DNA fragments called contiguous sequences, or "contigs." The contigs are then subjected to DNA sequencing, after which bioinformatics-based programs are used to arrange the contigs in their correct order on the basis of short overlapping sequences of nucleotides.

In this Exploring Genomics exercise you will carry out a simulation of contig alignment to help you understand the underlying logic of this approach to creating sequence maps of a chromosome. For this purpose, you will use the **National Center for Biotechnology Information BLAST** site and apply a DNA alignment program.

Exercise I – Arranging Contigs to Create a Chromosome Map

1. Access BLAST from the NCBI Web site at https://blast.ncbi.nlm.nih.gov/Blast.cgi. Locate and select the "Global Align" category under "Specialized searches." This feature allows you to compare two DNA sequences at a time to check for sequence similarity alignments.
2. Go to the Study Area, and open the Exploring Genomics exercise for this chapter. Listed are eight contig sequences, called Sequences A through H, taken from an actual human chromosome sequence deposited in GenBank. For this exercise we have used short fragments; however, in reality, contigs are usually several thousand base pairs long. To complete this exercise, copy one sequence into the "Enter Query Sequence" box and one sequence into the "Enter Subject Sequence" box and then run an alignment (by clicking on "Align"). Repeat these steps with other combinations of two sequences to determine which sequences overlap, and then use your findings to create a sequence map that places overlapping contigs in their proper order. Here are a few tips to consider:
   - Develop a strategy to be sure that you analyze alignments for all pairs of contigs.
   - Only consider alignment overlaps that show 100 percent sequence identity.
3. On the basis of your alignment results, answer the following questions, referring to the sequences by their letter codes (A through H):
   a. What is the correct order of overlapping contigs?
   b. What is the length, measured in number of nucleotides, of each sequence overlap between contigs?
   c. What is the total size of the chromosome segment that you assembled?
   d. Did you find any contigs that do not overlap with any of the others? Explain.
4. Run a nucleotide-nucleotide BLAST search with any one of the overlapping contigs to determine which chromosome these contigs were taken from, and report your answer.

CASE STUDY Your microbiome may be a risk factor for disease

A number of genes involved in susceptibility to inflammatory bowel disorders (IBDs), including Crohn's disease and ulcerative colitis, have been identified. However, it is clear that other nongenetic risk factors often trigger the onset of these diseases. As noted in Section 21.8, the Human Microbiome Project has provided valuable insights about the roles of the gut microbiome and its impact on intestinal disorders, including IBD. It is known that the gut microbiome of those with IBD is different from those in remission, and it is also different from individuals who do not have IBD. These observations suggest that transfer of gut microbiota from healthy individuals via fecal microbiota transplantation (FMT) might be a successful treatment for IBD. This idea is supported by the use of FMT for a potentially life-threatening form of colitis caused by the bacterium *Clostridium difficile*. After successful clinical trials, the U.S. Federal Drug Administration (FDA) has classified FMT as an investigational new drug. However, until it is formally approved, FMT can only be used to treat *C. difficile* infections that are resistant to antibiotic therapy.

1. If you had Crohn's disease or ulcerative colitis and wanted to undertake FMT, what genetic analyses might you consider to inform yourself about human genes, microbial genes, or the constitution of your gut microbiota and their correlation to or roles in Crohn's or ulcerative colitis before undergoing FMT?
2. The use of FMT, whether in a physician's office or at home, raises a number of ethical issues. What might they be, and which of them would concern you the most?
3. Several Internet sources offer screened donor fecal samples for use in FMT. What risks would you assume in undertaking this therapy at home using these samples? If you are willing to use this therapy on yourself, would you use it on one of your children?

For related reading, see Daloiso, V., et al. (2015). Ethical aspects of fecal microbiota transplantation (FMT). *Eur Rev Med Pharmacol Sci* 19(17):3173–3180.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. High-throughput computer-automated DNA sequencing methods coupled with bioinformatics enable scientists to assemble sequence maps of entire genomes.
2. Bioinformatics is essential for the analysis of genomes, transcriptomes, and proteomes. Bioinformatics applies computer hardware and software together with statistical approaches to analyze biological sequence data.
3. Annotation is used to identify protein-coding DNA sequencing and noncoding sequences such as regulatory elements, while bioinformatics programs are used to identify open reading frames that predict possible polypeptides coded for by a particular sequence.
4. Functional genomics predicts gene function based on sequence analysis.
5. The Human Genome Project revealed many surprises about human genetics, including gene number, the high degree of DNA sequence similarity between individuals and between humans and other species, and showed that many genes encode multiple proteins.
6. Genomics has led to other related "omics" disciplines that are rapidly changing how modern biologists study DNA, RNA, and proteins and many aspects of cell function.
7. The genomes for many important model organisms have been completed. Genomic analysis of model bacteria and eukaryotes has revealed similarities and important fundamental differences in genome size, gene number, and genome organization.
8. Increasingly, the importance of genomic variation is emerging through personalized genomics and other approaches, thus redefining the concept of a reference genome.
9. Studies in comparative genomics are revealing fascinating similarities and differences in genomes from different organisms, including the identification and analysis of gene families.
10. Metagenomics, or environmental genomics, sequences genomes of microorganisms from environmental samples, often identifying new sequences that encode proteins with novel functions.
11. DNA microarrays are valuable for transcriptome analysis in studying RNA expression patterns for thousands of genes simultaneously. Increasingly RNA sequencing (RNA-seq) is demonstrating its value for transcriptome analysis.
12. Methods such as two-dimensional gel electrophoresis and mass spectrometry are valuable for analyzing proteomes—the protein content of a cell.

INSIGHTS AND SOLUTIONS

1. One of the main problems in annotation is deciding how long a putative ORF must be before it is accepted as a gene. Shown on p. 519 are three different ORF scans of the same *E. coli* genome region—the region containing the *lacY* gene. Regions shaded in brown indicate ORFs. The top scan was set to accept ORFs of 50 nucleotides as genes. The middle and bottom scans accepted ORFs of 100 and 300 nucleotides as genes, respectively. How many putative genes are detected in each scan? The longest ORF covers 1254 bp; the next longest, 234 bp; and the shortest, 54 bp. How can we decide the actual number of genes in this region? In this type of ORF scan, is it more likely that the number of genes in the genome will be overestimated or underestimated? Why?

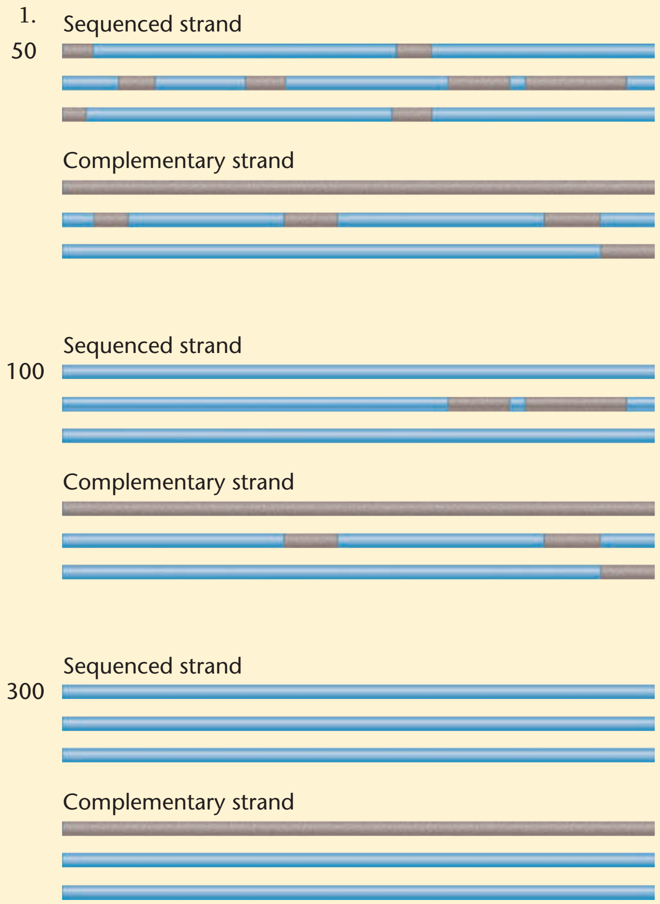
Solution:

Generally, one can examine conserved sequences in other organisms to indicate that an ORF is likely a coding region. One can also match a sequence to previously described sequences that are known to code for proteins. The problem is not easily solved—that is, deciding which ORF is actually a gene. The shorter the ORFs scan, the more likely the overestimate of genes because ORFs longer than 200 are less likely to occur by chance. For these scans, notice that the 50-bp scans produce the highest number of possible genes, whereas the 300-bp scan produces the lowest number (1) of possible genes.

2. Sequencing of the heterochromatic regions (repeat-rich sequences concentrated in centromeres and telomeres) of the *Drosophila* genome indicates that within 20.7 Mb, there are 297 protein-coding genes (Misra et al. 2002, *Genome Biology*, 3:research0083.1). Given that the euchromatic regions of the genome contain 13,379 protein-coding genes in 116.8 Mb, what general conclusion is apparent?

Solution:

Gene density in euchromatic regions of the *Drosophila* genome is about one gene per 8730 base pairs, while gene density in heterochromatic regions is one gene per 70,000 bases (20.7 Mb/297). Clearly, a given region of heterochromatin is much less likely to contain a gene than the same-sized region in euchromatin.



Problems and Discussion Questions

1. **HOW DO WE KNOW?** In this chapter, we focused on the analysis of genomes, transcriptomes, and proteomes and considered important applications and findings from these endeavors. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) How do we know which contigs are part of the same chromosome?
(b) How do we know if a genomic DNA sequence contains a protein-coding gene?
(c) What evidence supports the concept that humans share substantial sequence similarities and gene functional similarities with model organisms?
(d) How can proteomics identify differences between the number of protein-coding genes predicted for a genome and the number of proteins expressed by a genome?
(e) How has the concept of a reference genome evolved to encompass a broader understanding of genomic variation in humans?
(f) How have microarrays demonstrated that, although all cells of an organism have the same genome, some genes are expressed in almost all cells, whereas other genes show cell- and tissue-specific expression?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 485. All of these pertain to how genomics, bioinformatics, and proteomics approaches have changed how scientists study genes and proteins. Write a short essay that explains how recombinant DNA techniques were used to identify and study genes compared to how modern genomic techniques have revolutionized the cloning and analysis of genes.

3. What is functional genomics? How does it differ from comparative genomics?

4. Compare and contrast WGS to a map-based cloning approach.

5. What is bioinformatics, and why is this discipline essential for studying genomes? Provide two examples of bioinformatics applications.

6. Annotation involves identifying genes and gene-regulatory sequences in a genome. List and describe characteristics of a genome that are hallmarks for identifying genes in an unknown sequence. What characteristics would you look for in a bacterial genome? A eukaryotic genome?

7. How do high-throughput techniques such as computer-automated, next-generation sequencing, and mass spectrometry facilitate research in genomics and proteomics? Explain.

8. BLAST searches and related applications are essential for analyzing gene and protein sequences. Define BLAST, describe basic features of this bioinformatics tool, and give an example of information provided by a BLAST search.

9. What functional information about a genome can be determined through applications of chromatin immunoprecipitation (ChIP)?

10. Describe three major goals of the Human Genome Project.

11. Describe the human genome in terms of genome size, the percentage of the genome that codes for proteins, how much is composed of repetitive sequences, and how many genes it contains. Describe two other features of the human genome.

12. The Human Genome Project has demonstrated that in humans of all races and nationalities approximately 99.9 percent of the genome sequence is the same, yet different individuals can be identified by DNA fingerprinting techniques. What is one primary variation in the human genome that can be used to distinguish different individuals? Briefly explain your answer.

13. Through the Human Genome Project (HGP), a relatively accurate human genome sequence was published from combined samples from multiple individuals. It serves as a reference for a haploid genome. How do results from personal genome projects (PGP) differ from those of the HGP?

14. Explain differences between whole-genome sequencing (WGS) and whole-exome sequencing (WES), and describe advantages and disadvantages of each approach for identifying disease-causing mutations in a genome. Which approach was used for the Human Genome Project?

15. Describe the significance of the Genome 10K project.

16. It can be said that modern biology is experiencing an "omics" revolution. What does this mean? Explain your answer.

17. Metagenomics studies generate very large amounts of sequence data. Provide examples of genetic insight that can be learned from metagenomics.

18. What are DNA microarrays? How are they used?

19. Annotation of the human genome sequence reveals a discrepancy between the number of protein-coding genes and the number of predicted proteins actually expressed by the genome. Proteomic analysis indicates that human cells are capable of synthesizing more than 100,000 different proteins and perhaps three times this number. What is the discrepancy, and how can it be reconciled?

20. In Section 21.10 we briefly discussed the Human Proteome Map (HPM). An interactive Web site for the HPM is available at http://www.humanproteomemap.org. Visit this site, and then answer the questions in parts (a) and (b) and complete part (c).

(a) How many proteins were identified in this project?

(b) How many fetal tissues were analyzed?

(c) Use the "Query" tab and select the "Gene family" dropdown menu to do a search on the distribution of proteins encoded by a pathway of interest to you. Search in fetal tissues, adult tissues, or both.

Extra-Spicy Problems

21. Researchers have compared candidate loci in humans and rats in search of loci in the human genome that are likely to contribute to the constellation of factors leading to hypertension [Stoll, M., et al. (2000). *Genome Res.* 10:473–482]. Through this research, they identified 26 chromosomal regions that they consider likely to contain hypertension genes. How can comparative genomics aid in the identification of genes responsible for such a complex human disease? The researchers state that comparisons of rat and human candidate loci to those in the mouse may help validate their studies. Why might this be so?

22. Homology can be defined as the presence of common structures because of shared ancestry. Homology can involve genes, proteins, or anatomical structures. As a result of "descent with modification," many homologous structures have adapted different purposes.

(a) List three anatomical structures in vertebrates that are homologous but have different functions.

(b) Is it likely that homologous proteins from different species have the same or similar functions? Explain.

(c) Under what circumstances might one expect proteins of similar function to not share homology? Would you expect such proteins to be homologous at the level of DNA sequences?

23. Comparisons between human and chimpanzee genomes indicate that a gene that may function as a wild-type or normal gene in one primate may function as a disease-causing gene in another [The Chimpanzee Sequencing and Analysis Consortium (2005). *Nature* 437:69–87]. For instance, the *PPARG* locus (regulator of adipocyte differentiation) is a wild-type allele in chimps but is clearly associated with Type 2 diabetes in humans. What factors might cause this apparent contradiction? Would you consider such apparent contradictions to be rare or common? What impact might such findings have on the use of comparative genomics to identify and design therapies for disease-causing genes in humans?

24. Genomic sequencing has opened the door to numerous studies that help us understand the evolutionary forces shaping the genetic makeup of organisms. Using databases containing the sequences of 25 genomes, scientists examined the relationship between GC content and global amino acid composition [Kreil, D. P., and Ouzounis, C. A. (2001) *Nucl. Acids Res.* 29:1608–1615]. They found that it is possible to identify thermophilic species on the basis of their amino acid composition alone, which suggests that evolution in a hot environment selects for a certain whole organism amino acid composition. In what way might evolution in extreme environments influence genome and amino acid composition? How might evolution in extreme environments influence the interpretation of genome sequence data?

25. Whole-exome sequencing (WES) is helping physicians diagnose a genetic condition that has defied diagnosis by traditional means. The implication here is that exons in the nuclear genome are sequenced in the hopes that, by comparison with the genomes of nonaffected individuals, a diagnosis might be revealed.

(a) What are the strengths and weaknesses of this approach?

(b) If you were ordering WES for a patient, would you also include an analysis of the patient's mitochondrial genome?

26. Recall that when the HGP was completed, more than 40 percent of the genes identified had unknown functions. The PANTHER database provides access to comprehensive and current functional assignments for human genes (and genes from other species).

Go to http://www.pantherdb.org/data/. In the frame on the left side of the screen locate the "Quick links" and use the "Whole genome function views" link to a view of a pie chart of current functional classes for human genes. Mouse over the pie chart to answer these questions. What percentage of human genes encode transcription factors? Cytoskeletal proteins? Transmembrane receptor regulatory/adaptor proteins?

22

Applications of Genetic Engineering and Biotechnology

The *E. coli* in these colonies have been genetically engineered to produce lycopene, an antioxidant found in tomatoes, which also gives the fruit its characteristic color.

CHAPTER CONCEPTS

- Recombinant DNA technology, genetic engineering, and biotechnology have revolutionized medicine and agriculture.
- Genetically modified organisms, including transgenic animals, can serve as bioreactors to produce therapeutic proteins as biopharmaceutical products.
- Genetic modifications of plants have resulted in herbicide- and pest-resistant crops, and crops with improved nutritional value.
- Applications of recombinant DNA technology and genomics have become essential for diagnosing genetic disorders.
- Whole-genome sequencing (WGS) of an individual's DNA is increasingly being used for genetic testing.
- Genome-wide association studies (GWAS) scan for hundreds or thousands of genetic differences in an attempt to link genome variations to particular traits and diseases.
- Functional synthetic genomes have been assembled, elevating interest in potential applications of synthetic biology.
- Almost all applications of genetic engineering and biotechnology present unresolved ethical dilemmas that involve important moral, social, and legal issues.

Earlier in the text (see Chapters 20 and 21), we reviewed the field of recombinant DNA technology and its role in genomic analysis. Thus, we have established the many ways that the genetic material can be manipulated to not only expand our knowledge of the genetic constitution of organisms, but to genetically engineer them. In this chapter, we turn to a consideration of the many applications that these technologies have made possible.

Genetic engineering refers to the alteration of an organism's genome and typically involves the use of recombinant DNA technologies to add a gene or genes to a genome, but it can also involve gene removal. The ability to manipulate DNA *in vitro* and to introduce genes into living cells has allowed scientists to generate new varieties of plants, animals, and other organisms with specific traits. These organisms are called **genetically modified organisms (GMOs).** Industry analysts estimate that genetic engineering will, in the near future, lead to U.S. commercial products worth over $60 billion. Many of these products will be developed by the **biotechnology** industry.

Biotechnology uses living organisms to create products or processes that help improve the quality of life for humans or other organisms. Biotechnology as a modern industry began in earnest shortly after recombinant DNA technology developed. But biotechnology is actually a science dating back to ancient civilization and the use of microbes to make many important products, including beverages such as wine and beer, vinegar, breads, and cheeses. Modern biotechnology relies heavily on genetic engineering and genomics applications, and these areas will be highlighted in this chapter. Existing products and new developments that occur seemingly every day make the biotechnology industry one of the most rapidly developing branches of the workforce worldwide, encompassing nearly 5000 companies in 54 countries.

This chapter will present only a selection of applications that illustrate the power of genetic engineering and biotechnology and the dilemmas they engender. We briefly describe how genetic engineering has affected the production of pharmaceutical products and consider examples of genetic engineering in animals. We examine the impact of different genetic technologies on the diagnosis of human diseases and explore the concept of synthetic genomes. Finally, we consider some of the social, ethical, and legal implications of genetic engineering and biotechnology.

Please note that many of the topics discussed in this chapter are covered in more detail later in the text (see the Special Topic chapters found at the end of the book).

22.1 Genetically Engineered Organisms Synthesize a Variety of Valuable Biopharmaceutical Products

The most successful and widespread application of recombinant DNA technology by the biotechnology industry has been production of recombinant proteins as **biopharmaceutical** products—particularly, therapeutic proteins to treat diseases. Prior to the recombinant DNA era, biopharmaceuticals such as insulin, clotting factors, or growth hormones were purified from the pancreas, blood, or pituitary glands, respectively. These tissue sources were in limited supply, and the purification processes were expensive. In addition, products derived from these natural sources could be contaminated by disease agents such as viruses. Once human genes encoding important therapeutic proteins could be cloned and expressed in a number of host-cell types, we had more abundant, safer, and less expensive sources of biopharmaceuticals. **Biopharming** is a commonly used term to describe the production of valuable proteins in genetically modified (GM) animals and plants.

Here we outline several examples of therapeutic products produced by expression of cloned genes in transgenic host cells and organisms. It should not surprise you that cancers, arthritis, diabetes, heart disease, and infectious diseases such as AIDS are among the major diseases that biotechnology companies are targeting for treatment by recombinant therapeutic products. **Table 22.1** provides a short list of important recombinant products currently synthesized in transgenic bacteria, plants, yeast, and animals.

Recombinant Protein Production in Bacteria

Many therapeutic proteins have been produced by introducing human genes into bacteria. In most cases, the human gene is cloned into a plasmid, and the recombinant vector is introduced into the bacterial host. Large quantities of the transformed bacteria are grown, and the recombinant human protein is recovered and purified from bacterial extracts.

The first human gene product manufactured by recombinant DNA technology was human insulin, called Humulin®, which was licensed for therapeutic use in 1982 by the **U.S. Food and Drug Administration (FDA),** the government agency responsible for regulating the safety of

TABLE 22.1 Examples of Genetically Engineered Biopharmaceutical Products Available or Under Development

Gene Product	Condition Treated	Host Type
Erythropoietin	Anemia	*E. coli*, cultured mammalian cells
Interferons	Multiple sclerosis, cancer	*E. coli*, cultured mammalian cells
Tissue plasminogen activator (tPA)	Heart attack, stroke	Cultured mammalian cells
Human growth hormone	Dwarfism	Cultured mammalian cells
Sebelipase alfa	Lysosomal acid lipase deficiency	Transgenic chickens
Human clotting factor VIII	Hemophilia A	Transgenic sheep, pigs
C1 inhibitor	Hereditary angioedema	Transgenic rabbits
Recombinant human antithrombin	Hereditary antithrombin deficiency	Transgenic goats
Hepatitis B surface protein vaccine	Hepatitis B infections	Cultured yeast cells, bananas
Immunoglobulin IgG1 to HSV-2	Herpesvirus infections	Transgenic soybeans
Recombinant monoclonal antibodies	Passive immunization against rabies, cancer, rheumatoid arthritis	Transgenic tobacco, soybeans, cultured mammalian cells
Ebola antibodies	Ebola virus	Transgenic tobacco

food and drug products and medical devices. In 1977, scientists at Genentech, a San Francisco biotechnology company cofounded in 1976 by Herbert Boyer (one of the pioneers of using plasmids for recombinant DNA technology) and Robert Swanson, isolated and cloned the gene for insulin and expressed it in bacterial cells. Genentech, short for "genetic engineering technology," is also generally regarded as the world's first biotechnology company.

Previously, insulin was chemically extracted from the pancreas of cows and pigs obtained from slaughterhouses. **Insulin** is a protein hormone that regulates glucose metabolism. Individuals who cannot produce insulin have diabetes, a disease that, in its more severe form (Type 1), affects more than 1.25 million individuals in the United States. Although synthetic human insulin can now be produced by another process, a look at the original genetic engineering method is instructive, as it shows both the promise and the difficulty of applying recombinant DNA technology.

Clusters of specialized cells in the pancreas synthesize a precursor polypeptide known as preproinsulin. As this polypeptide is secreted from the cell, amino acids are cleaved from the end and the middle of the chain. These cleavages produce the mature insulin molecule, which consists of two short polypeptide chains (the A and B chains) joined by disulfide bonds. The A subunit contains 21 amino acids, and the B subunit contains 30.

In the original bioengineering process, synthetic genes that encode the A and B subunits were constructed by oligonucleotide synthesis (63 nucleotides for the A polypeptide and 90 nucleotides for the B polypeptide). Each synthetic oligonucleotide was inserted into a separate vector, adjacent to the *lacZ* gene encoding the bacterial form of the enzyme β-galactosidase. When transferred to a bacterial host, the *lacZ* gene and the adjacent synthetic oligonucleotide were transcribed and translated as a unit. The product is a **fusion protein**—that is, a hybrid protein consisting of the amino acid sequence for β-galactosidase attached to the amino acid sequence for one of the insulin subunits (**Figure 22.1**). The fusion proteins were purified from bacterial extracts and treated with cyanogen bromide, a chemical that cleaves the insulin subunits from the β-galactosidase. When mixed, the two insulin subunits spontaneously united, forming an intact, active insulin molecule. The purified injectable insulin was then packaged for use by diabetics.

Shortly after insulin became available, growth hormone—used to treat children who suffer from a form of dwarfism—was cloned. Soon, recombinant DNA technology made that product readily available too, as well as a wide variety of other medically important proteins that were once difficult to obtain in adequate amounts. Since recombinant insulin ushered in the biotechnology era, well over 200 recombinant biopharmaceutical products have entered

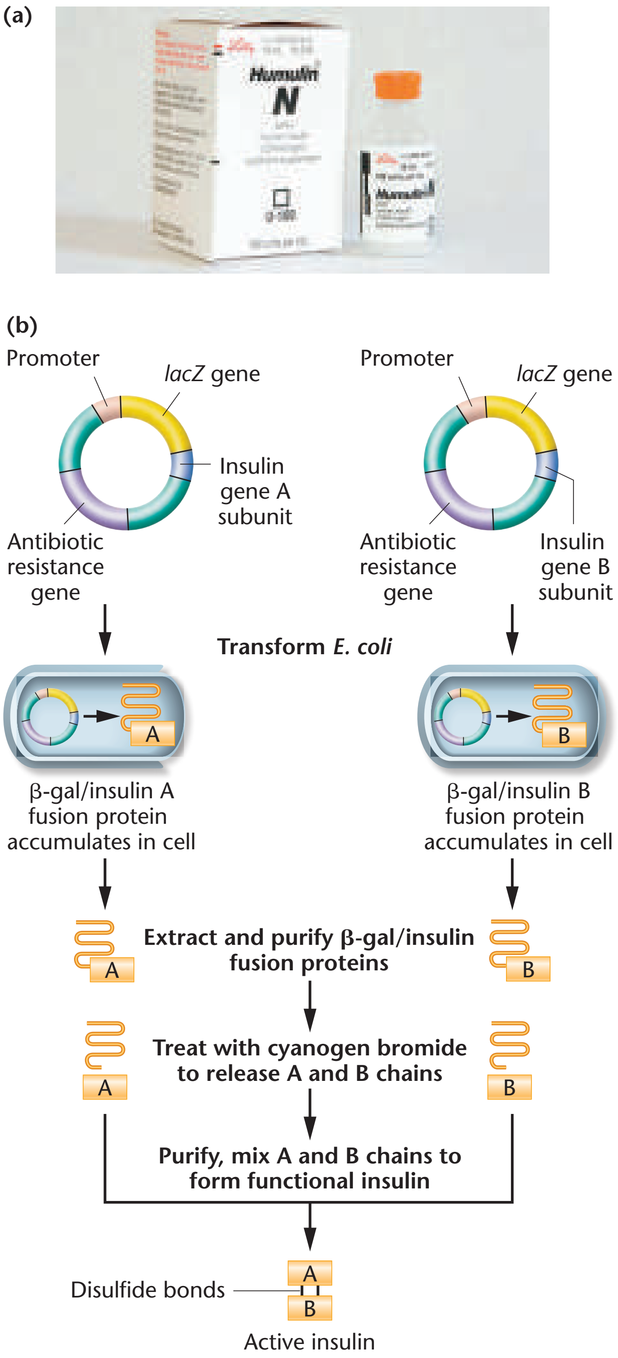

(a)
Humulin N
(b)
Promoter
lacZ gene
Insulin gene A subunit
Antibiotic resistance gene
Promoter
lacZ gene
Antibiotic resistance gene
Insulin gene B subunit
Transform *E. coli*
A
B
β-gal/insulin A fusion protein accumulates in cell
β-gal/insulin B fusion protein accumulates in cell
Extract and purify β-gal/insulin fusion proteins
A
B
Treat with cyanogen bromide to release A and B chains
A
B
Purify, mix A and B chains to form functional insulin
Disulfide bonds
A
B
Active insulin

FIGURE 22.1 (a) Humulin, a recombinant form of human insulin, was the first therapeutic protein produced by recombinant DNA technology to be approved for use in humans. (b) To synthesize recombinant human insulin, synthetic oligonucleotides encoding the insulin A and B chains were inserted (in separate vectors) adjacent to the *E. coli lacZ* gene. Recombinant plasmids were transformed into *E. coli* host cells, where the β-gal/insulin fusion protein was synthesized and accumulated in the cells. Fusion proteins were then extracted from the host cells and purified. Insulin chains were released from β-galactosidase by treatment with cyanogen bromide. The insulin subunits were purified and mixed to produce a functional insulin molecule.

the market worldwide. In recent years, development of many other, nonbiopharmaceutical products has been a very active area of research. One example includes the production in *E. coli* of the antioxidant lycopene found in tomatoes (see the opening photograph at the beginning of this chapter). Lycopene produced by *E. coli* is not yet available for human consumption.

Transgenic Animal Hosts and Biopharmaceutical Products

Although bacteria are widely used to produce therapeutic proteins, there are some disadvantages in using bacterial hosts to synthesize eukaryotic proteins. One problem is that bacterial cells often cannot process and modify eukaryotic proteins. As a result, they frequently cannot add the carbohydrates and phosphate groups to proteins that are needed for full biological activity. In addition, eukaryotic proteins produced in bacterial cells often do not fold into the proper three-dimensional conformation and are therefore inactive. To overcome these difficulties and increase yields, many biopharmaceuticals are now produced in eukaryotic hosts.

Yeast are also valuable hosts for expressing therapeutic proteins. Because they are eukaryotes, many of the mechanisms yeast use for protein processing are similar to human cells, and this can be an advantage over bacterial cells for producing functional recombinant proteins. Even insect cells are valuable for producing recombinant proteins, through the use of a viral gene delivery system called baculovirus. Recombinant baculovirus containing a gene of interest is used to infect insect cell lines, which then express the protein at high levels. Baculovirus-insect cell expression is particularly useful for producing human recombinant proteins that are heavily glycosylated. As seen in Table 22.1, eukaryotic hosts may include cultured eukaryotic cells (from both plant or animal sources) or transgenic farm animals. For example, a herd of goats or cows can serve as very effective **bioreactors** or **biofactories**—living factories—that could continuously make milk containing the desired therapeutic protein that can then be isolated in a noninvasive way.

Regardless of the host, therapeutic proteins may then be purified from the host cells—or when **transgenic animals** are used, isolated from animal products such as milk. (Refer to our earlier discussion in Chapter 20 on how transgenic animals can be created to allow for expression of a transgene of interest.)

In 2006, recombinant human **antithrombin,** an anticlotting protein, became the world's first drug extracted from the milk of farm animals to be approved for use in humans. Scientists at GTC Biotherapeutics introduced the human antithrombin gene into goats. By placing the gene adjacent to a promoter for beta casein, a common protein in milk, GTC scientists were able to target antithrombin expression in the mammary gland. As a result, antithrombin protein is abundantly present in the milk of these goats. In one year, a single goat will produce the equivalent amount of antithrombin that in the past would have required ~90,000 human blood collections.

In 2015, the FDA approved only the fourth recombinant protein drug expressed from a transgenic animal—**sebelipase alfa,** a recombinant enzyme purified from the egg whites of transgenic hens (*Gallus gallus*). Sebelipase alfa is approved for treating severe or fatal forms of a lysosomal enzyme deficiency (lysosomal acid lipase), including a condition that can cause liver fibrosis, cirrhosis, and liver failure. A transgenic chicken bioreactor approach was used because it results in a glycosylation pattern for sebelipase alfa that is essential for biological activity of the enzyme.

Recombinant DNA Approaches for Vaccine Production

Another successful application of recombinant DNA technology for therapeutic purposes is the production of vaccines. Vaccines stimulate the immune system to produce antibodies against disease-causing organisms and thereby confer immunity against specific diseases. Traditionally, two types of vaccines have been used: **inactivated vaccines,** which are prepared from killed samples of the infectious virus or bacteria; and **attenuated vaccines,** which are live viruses or bacteria that can no longer reproduce but can cause a mild form of the disease. Inactivated vaccines include the vaccines for rabies and influenza; vaccines for tuberculosis, cholera, and chickenpox are examples of attenuated vaccines.

Genetic engineering is being used to produce **subunit vaccines,** which consist of one or more surface proteins from the virus or bacterium rather than the entire virus or bacterium. This surface protein acts as an antigen that stimulates the immune system to make antibodies that act against the organism from which it was derived. Often the surface protein is produced through recombinant DNA technology by cloning and expressing the genes encoding the protein to be used for the vaccine. One of the first subunit vaccines was made against the **hepatitis B virus,** which causes liver damage and cancer. The gene that encodes the hepatitis B surface protein was cloned into a yeast expression vector, and the cloned gene was expressed in yeast host cells. The protein was then extracted and purified from the host cells and packaged for use as a vaccine.

In 2005, the FDA approved **Gardasil®,** a subunit vaccine produced by the pharmaceutical company Merck and the first cancer vaccine to receive FDA approval. Gardasil targets four strains of **human papillomavirus (HPV)** that cause ~70 percent of cervical cancers. Approximately

70 percent of sexually active women will be infected by an HPV strain during their lifetime. Gardasil is designed to provide immune protection against HPV prior to infection but is not effective against existing infections. You may have heard of Gardasil through media coverage of the legislation proposed in several states that would require all adolescent schoolgirls to receive a Gardasil vaccination regardless of whether or not they are sexually active.

Vaccine Proteins Can Be Produced by Plants

Plants offer several other advantages for expressing recombinant proteins. For instance, once a transgenic plant is created, it can easily be grown and replicated in a greenhouse or field, and it will provide a constant source of recombinant protein. In addition, the cost of expressing a recombinant protein in a transgenic plant is typically much lower than making the same protein in bacteria, yeast, or mammalian cells.

No recombinant proteins expressed in transgenic plants have yet been approved for use by the FDA as therapeutic proteins for humans, although about a dozen products are in their final clinical trials.

In 2014, an outbreak of Ebola virus in West Africa killed over 1500 people, with many more cases likely unreported. Ebola causes hemorrhagic fever and produces fatality rates of approximately 90 percent. There is currently no effective treatment for curing or preventing Ebola virus infection. But antibodies against Ebola expressed in tobacco leaves are showing promise in ongoing clinical trials. Mice were used to create monoclonal antibodies against the virus. The antibody genes were then introduced into tobacco plants. The transgenic tobacco plants express high quantities of the antibody proteins, which can then be isolated and purified for use in humans. Transgenic tobacco plants are commonly used for expressing recombinant proteins because of the large size of their leaves and relatively high yield of recombinant proteins compared to other plants.

In spite of the promise of new vaccines, developing countries face serious difficulties in manufacturing, transporting, and storing them. Most vaccines need refrigeration and must be injected under sterile conditions. In many rural areas, refrigeration and sterilization facilities are not available. In addition, in many cultures people are fearful of being injected with needles. To overcome these problems, scientists are attempting to develop vaccines that can be synthesized in edible food plants. These vaccines would need to be inexpensive to produce, would not require refrigeration, and would not have to be given under sterile conditions by trained medical personnel. Some edible vaccines are now in clinical trials. For example, a vaccine against a bacterium that causes cholera has been produced in genetically engineered potatoes and used to successfully vaccinate human volunteers.

Trials using vaccine-producing bananas are currently under way. Bananas are considered to be perhaps the best edible vaccine candidate for a hepatitis B vaccine, in part because they don’t have to be cooked before being eaten. But several technical questions about vaccine delivery in plants need to be answered if edible plant vaccines are to become more widely used. For example, how can vaccine dose be carefully controlled when fruits and vegetables grow to different sizes and express different amounts of the vaccine? Will vaccine proteins pass through the digestive tract unaltered so that they maintain their ability to provide immune protection? Nevertheless, these are exciting prospects.

DNA-Based Vaccines

Lastly, it warrants mentioning that **DNA-based vaccines** have been attempted for many years, and recently there has been renewed interest in using such vaccines to protect against viral pathogens. In this approach, DNA encoding proteins from a particular pathogen are inserted into plasmid vectors, which are then injected directly into an individual or delivered via a viral vector similar to the way certain viruses are used for gene therapy. The rationale behind this approach is that pathogen proteins encoded by the delivered DNA would be produced and trigger an immune response that could provide protection should an immunized person be exposed to the pathogen in the future.

For example, in 2015 an outbreak of the mosquito-borne Zika virus (ZIKV) in the Caribbean and the Americas was identified as the cause of congenital defects, such as microcephaly, in babies from women infected with ZIKV. Thus, a priority was to develop a vaccine to establish ZIKV immunity in women of childbearing age and pregnant women at risk for ZIKV infection that could harm a developing fetus. Early-stage trials with DNA expressing the premembrane and envelope proteins of ZIKV have been shown to create levels of antibodies in mice and Rhesus macaques capable of neutralizing the ZIKV and protecting against ZIKV infection. Clinical trials of the DNA vaccine for ZIKV have already begun, so it will be very interesting to see if this approach is effective in humans.

Similarly, several trials are under way using plasmid DNA encoding protein antigens from HIV. However, a major limitation of DNA-based vectors, evident in the HIV research, has been that they typically result in very low production of protein encoded by delivered genes, and thus the immune response in vaccinated persons is insufficient to provide the desired protection. Work on DNA-based vaccines continues to be an active area of exploration, but whether they will ever have significant roles in the vaccine market remains to be seen.

NOW SOLVE THIS

22.1 In order to vaccinate people against diseases by having them eat antigens (such as the cholera toxin) or antibodies expressed in an edible vaccine, the antigen must reach the cells of the small intestine in order to enter the bloodstream. What are some potential problems of this method?

- **HINT:** *This problem asks you to consider why edible vaccines may not be effective. The key to its solution is to consider the molecular structure of the antigen or antibody and its recognition by the immune system.*

22.2 Genetic Engineering of Plants Has Revolutionized Agriculture

For millennia, farmers have manipulated the genetic makeup of plants and animals to enhance food production. Until the advent of genetic engineering 30 years ago, these genetic manipulations were primarily restricted to **selective breeding**—the selection and breeding of naturally occurring or mutagen-induced variants. In the last 50 to 100 years, genetic improvement of crop plants through the traditional methods of artificial selection and genetic crosses has resulted in dramatic increases in productivity and nutritional enhancement. For example, maize yields have increased fourfold over the last 60 years, and more than half of this increase is due to genetic improvement by artificial selection and selective breeding (**Figure 22.2**). Modern maize has substantially larger ears and kernels than the predecessor crops, including hybrids from which it was bred.

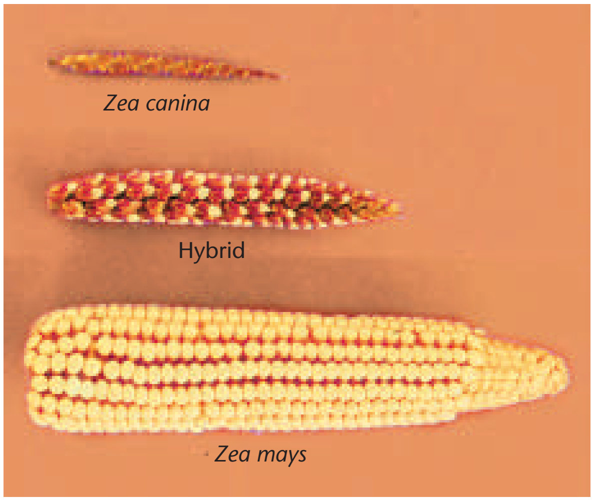


FIGURE 22.2 Selective breeding is one of the oldest methods of genetic alteration of plants. Shown here is teosinte (*Zea canina*, top), a selectively bred hybrid (center), and modern corn (*Zea mays*, bottom).

Since the advent of recombinant DNA technology, it became possible to identify, isolate, and clone genes that confer desired traits in plants and then use these genes to create transgenic plants. As a result, it is possible to create plants with enhanced qualities such as insect resistance, herbicide resistance, or nutritional characteristics—a primary purpose of **agricultural biotechnology.** (Genetic modifications of plant crops are discussed in greater detail in Special Topic Chapter 4—Genetically Modified Foods).

Worldwide, millions of acres of genetically engineered crops have been planted, particularly herbicide- and pest-resistant soybeans, corn, cotton, and canola; over 50 different transgenic crop varieties are available, including alfalfa, corn, rice, potatoes, tomatoes, tobacco, wheat, and cranberries.

Several of the main reasons for generating transgenic crops include:

- Improving the growth characteristics and yield of agriculturally valuable crops
- Increasing the nutritional value of crops
- Providing crop resistance against herbicides (the most widely used application of genetically modified crop plants), insects (insect resistance is the second leading application), viruses, and drought

In addition, many new GM crops that will soon be on the market will be designed for ethanol production and for making biodiesel fuel—that is, for providing sustainable sources of energy.

Insights from plant genome sequencing projects will undoubtedly be the catalyst for analysis of genetic diversity in crop plants, identification of genes involved in crop domestication and breeding traits, and subsequent enhancement of a variety of desirable traits through genetic engineering. In the past several years genome projects have been completed for many major food and industrial crops, including the three crops that account for most of the world's caloric intake: maize, rice, and wheat. The genome for a popular crop species of coffee plants was recently sequenced. Plant scientists expect to use genome data to improve coffee crop growth and eventually to improve crop phenotypes to produce the most desirable attributes for coffee beans.

22.3 Genetically Modified Animals Serve Important Roles in Biotechnology

Although genetically engineered plants are major players in modern agriculture, commercial applications of GM animals are less widespread. Nonetheless, some high-profile examples have aroused public interest and controversy.

Examples of Transgenic Animals

Most transgenic animals are created for research purposes to study gene function. For instance, mice containing a human growth hormone transgene were some of the first transgenic animals created. However, it is expected that transgenic animals created for commercial purposes may increasingly be available in the future.

Attempts to create farm animals containing transgenic growth hormone genes have not been particularly successful, probably because growth is a complex, multigene trait. One notable exception is the transgenic Atlantic salmon, bearing copies of a Chinook salmon growth hormone gene adjacent to a constitutive promoter (see Special Topic Chapter 4—Genetically Modified Foods). In 2015, after more than 20 years of regulatory review, the U.S. FDA approved these salmon as the first genetically modified animals acceptable for human consumption.

As discussed in Section 22.1, currently, the major uses for transgenic farm animals are as bioreactors to produce useful pharmaceutical products. Significant research efforts are also being made to protect farm animals against common pathogens that cause disease and animal loss (including potential bioweapon pathogens that could be used in a terrorist attack on food animals) and put the food supply at risk.

For instance, controlling mastitis in cattle by creating transgenic cows has shown promise. **Mastitis** is an infection of the mammary glands. It is the most costly disease affecting the dairy industry, leading to over $2 billion in losses in the United States. Mastitis can block milk ducts, reducing milk output, and can also contaminate the milk with pathogenic microbes. Infection by the bacterium *Staphylococcus aureus* is the most common cause of mastitis, and most cattle with mastitis typically do not respond well to conventional treatments with antibiotics. As a result, mastitis is a significant cause of herd reduction.

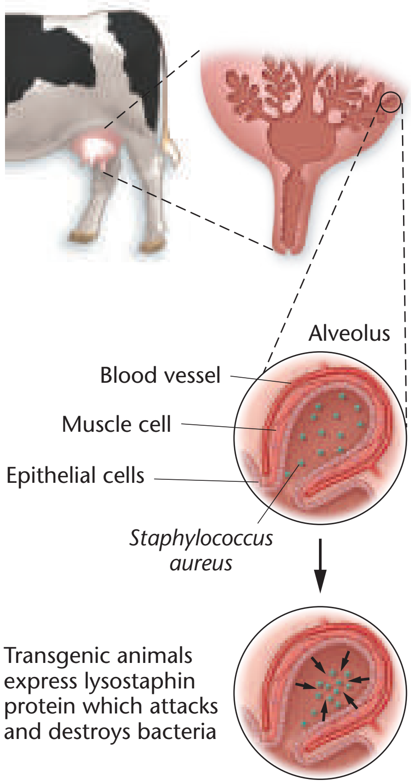

Alveolus
Blood vessel
Muscle cell
Epithelial cells
Staphylococcus aureus
Transgenic animals express lysostaphin protein which attacks and destroys bacteria

FIGURE 22.3 Transgenic cows for battling mastitis. The mammary glands of nontransgenic cows are highly susceptible to infection by the skin microbe *Staphylococcus aureus*. Transgenic cows express the lysostaphin transgene in milk, where it can kill *S. aureus* bacterium before they can multiply in sufficient numbers to cause inflammation and damage mammary tissue.

In an attempt to create cattle resistant to mastitis, transgenic cows were generated (Figure 22.3) that possessed the lysostaphin gene from *Staphylococcus simulans*. Lysostaphin is an enzyme that specifically cleaves components of the *S. aureus* cell wall. Transgenic cows expressing this protein in milk thus produce a natural antibiotic that wards off *S. aureus* infections. The mastitis problem is not completely solved in these transgenic cows, however, because lysostaphin is not effective against other microbes such as *E. coli* and *Streptococcus uberis* that occasionally cause mastitis; moreover, there is also the potential that *S. aureus* may develop resistance to lysostaphin. Nonetheless, scientists are cautiously optimistic that transgenic approaches have a strong future for providing farm animals with a level of protection against major pathogens.

Researchers in New Zealand have engineered a cow to produce hypoallergenic milk. This research effort has been spurred by the fact that an estimated 2–3 percent of babies are allergic to milk from dairy cows and develop a reaction to a protein called β-lactoglobulin. These researchers designed miRNAs to inhibit β-lactoglobulin expression and then used a transgenic approach to introduce genes for the miRNAs into cow embryos. Of 100 GM cow embryos, only one produced a calf, Daisy. As Daisy began lactating, researchers found that her milk did not have any detectable levels of β-lactoglobulin. Currently, studies are under way to determine if Daisy's milk is less allergenic to mice, with future plans to test whether humans are allergic to Daisy's milk.

Despite the fact that only one genetically engineered animal (transgenic salmon) has been approved for human consumption anywhere in the world, genetic engineering research on cattle, pigs and other animals continues. And improvements in gene editing approaches such as the CRISPR-Cas system (see Special Topic Chapter 1—CRISPR-Cas and Genome Editing) have accelerated some of these projects. For example, researchers have used CRISPR-Cas to disrupt the myostatin gene (*MSTN*) in cows and pigs. *MSTN* encodes a protein that normally inhibits growth of muscle cells to keep muscle growth proportional. When *MSTN* is disrupted by gene editing, muscle cells proliferate unchecked, creating "double-muscled" animals. Creators of these animals are hoping these animals can be used to increase meat yield per animal for making beef and pork products suitable for human consumption.

Yorktown Technologies of Austin, Texas, have even used genetic engineering to create a unique pet, called the **GloFish®**, a transgenic strain of zebrafish (*Danio rerio*) containing a red fluorescent protein gene from sea anemones. These fish fluoresce bright red when illuminated by ultraviolet light. Since developing these initial fish, Yorktown now offers fluorescent zebrafish in green, pink, yellow (**Figure 22.4**), and other colors, and has also created transgenic lines of GloFish tetras (*Paracheirodon axelrod*) based on the introduction of additional genes from anemones or jellyfish. GM critics describe these fish as an abuse of genetic technology. However, GloFish may not be as frivolous a use of genetic engineering as some believe. A variant of this transgenic model, incorporating a heavy-metal-inducible promoter adjacent to the red fluorescent protein gene, has shown promise in a bioassay for heavy-metal contamination of water. This was the original purpose behind the development of GloFish. When these transgenic zebrafish are in water contaminated by mercury and other heavy metals, the promoter becomes activated, inducing transcription of the red fluorescent protein gene. In this way, zebrafish fluorescence can be used as a bioassay to measure heavy metal contamination and uptake by living organisms.

We conclude this section by briefly discussing a recent application of a genetically modified mosquito. The public health threat of viral diseases transmitted by insects became particularly evident in 2016 with concern over the Zika virus and its possible spread to the United States. Specifically, the *Aedes aegypti* mosquito is considered a human health threat because it is one of the insect vectors that transmits Zika. Because there is currently no vaccine for Zika, and because *A. aegypti* is an invasive species that also transmits other pathogens such as dengue virus, scientists are attempting genetic engineering of the host mosquito to stop the spread of Zika virus.

One genetic engineering approach, which has also shown promise in the *Anopheles* mosquitos that transmit malaria, is called **gene drive.** Gene drive enables a particular gene (including a transgene or a gene that has been genetically modified) to be transmitted to a majority of an individual's offspring—rather than just to half of the offspring as would occur for most maternal or paternal copies of a gene on an autosome. In organisms that reproduce rapidly such as mosquitos, inserting a gene into just a few individuals by gene drive can quickly affect subsequent generations, thus forcing or "driving" the gene into the population.

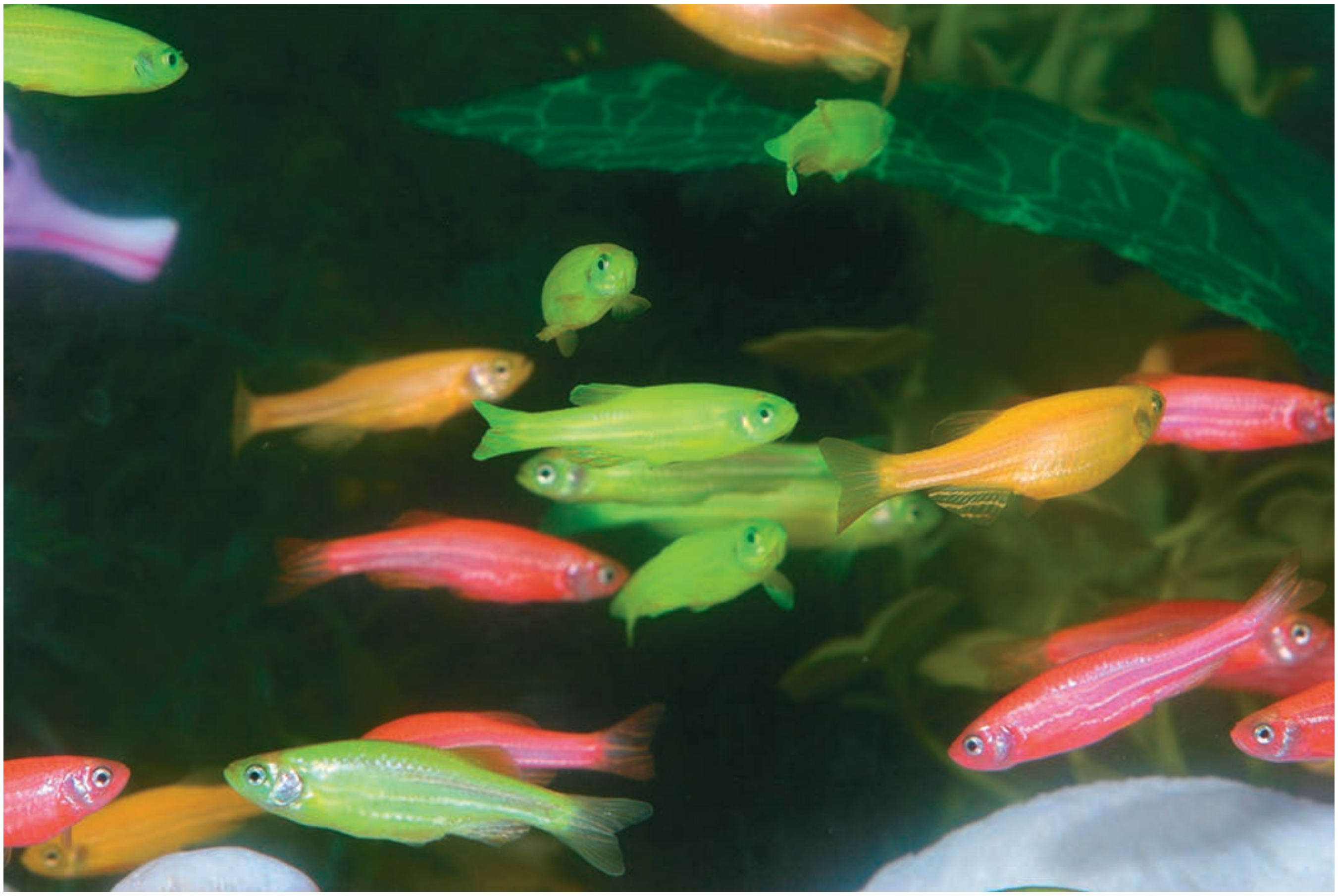


FIGURE 22.4 GloFish, marketed as the world's first GM pet, are a controversial product of genetic engineering.

A company based in the United Kingdom has created genetically modified *A. aegypti* male mosquitoes that do not bite or spread disease. But these mosquitoes mate with females and pass along a lethal gene that produces short-lived offspring that die before they reach adulthood. Field trials in the Cayman Islands, Panama, and Brazil have reduced local *A. aegypti* populations by 80–90 percent. Initially, the modified male mosquitoes were created by traditional transgenic approaches, but now researchers are using CRISPR-Cas for gene drive.

Because *A. aegypti* is viewed as a pest insect with no particular ecological value, the FDA has suggested that it prefers gene drive approaches instead of the widespread use of broad-spectrum insecticides, which kill other insects. But the use of gene drive for other potential applications is controversial because the technique can alter gene frequencies in populations. Gene drive is the topic of discussion in the Case Study at the end of this chapter. (You will also learn more about CRISPR-Cas approaches for gene drive applications in Special Topic Chapter 1—CRISPR-Cas and Genome Editing .)

22.4 Genetic Testing, Including Genomic Analysis, Is Transforming Medical Diagnosis

Gene-based technologies have had major impacts on the diagnosis of disease and are revolutionizing medical treatments based on the development of specific and effective pharmaceuticals. Because of the Human Genome Project and related advances in genomics, researchers have been making rapid progress in identifying genes involved in both single-gene diseases and complex genetic traits. In this section, we provide an overview of these developments.

Genetic Testing for Prognostic or Diagnostic Purposes

Genetic testing was one of the first successful applications of recombinant DNA technology, and currently more than 900 gene tests are in use. Increasingly, scientists and physicians can directly examine an individual's DNA for mutations associated with disease. These tests usually detect gene alterations associated with single-gene disorders inherited in a Mendelian fashion. Examples include sickle-cell anemia, cystic fibrosis, Huntington disease, hemophilia, and muscular dystrophy. Other genetic tests have been developed for complex disorders such as breast and colon cancers.

Gene tests are used for prenatal, childhood, and adult prognosis and diagnosis of genetic diseases; to identify carriers; and to identify genetic diseases in embryos created by *in vitro* fertilization. among other applications. For genetic testing of adults, DNA from white blood cells is commonly used. Alternatively, many genetic tests can be carried out on cheek cells, collected by swabbing the inside of the mouth, or on hair cells. Some genetic testing can be carried out on gametes.

What does it mean when a genetic test is performed for *prognostic* purposes, and how does this differ from a *diagnostic* test? A prognostic test predicts a person's likelihood of developing a particular genetic disorder. A diagnostic test for a genetic condition identifies a particular mutation or genetic change that causes the disease or condition. Sometimes a diagnostic test identifies a gene or mutation associated with a condition, but the test will not be able to determine whether the gene or mutation is the cause of the disorder or is a genetic variation that results from the condition.

Prenatal Genetic Testing

Although genetic testing of adults is increasing, over the past two decades more genetic testing has been used to detect genetic conditions in babies than in adults. In newborns, a simple prick of a baby's heel produces a few drops of blood that are used to check the newborn for many genetic disorders. In the United States, all states now require newborn screening for certain medical conditions (the number of diseases screened for is set by the individual state). There are currently about 60 conditions that can be detected, although many of these tests detect proteins or other metabolites and are not DNA- or RNA-based genetic tests.

Prenatal genetic tests, performed before a baby is born, are used for certain disorders in which waiting until birth is not desirable. For prenatal testing, fetal cells are obtained by **amniocentesis** or **chorionic villus sampling (CVS).** Figure 22.5 shows the procedure for amniocentesis, in which a small volume of the amniotic fluid surrounding the fetus is removed. Amniotic fluid contains fetal cells that can be used for karyotyping, genetic testing, and other procedures. For chorionic villus sampling, cells from the fetal portion of the placental wall (the chorionic villi) are sampled through a vacuum tube, and analyses can be carried out on this tissue. Captured fetal cells can then be subjected to genetic analysis by techniques that involve PCR (such as allele-specific oligonucleotide testing, described later in this section) or DNA sequencing.

Noninvasive procedures are also being developed for prenatal genetic testing of fetal DNA. These procedures reduce the risk to the fetus. Circulating in each person's bloodstream is DNA that is released from the person's dead and dying cells. This so-called *cell-free DNA (cfDNA)* is cut up into small fragments by enzymes in the blood. The blood of a pregnant woman also contains snippets of cfDNA from the fetus. It is estimated that ~3 to 6 percent of the DNA in a pregnant mother's blood belongs to her baby. It is now possible to analyze these traces of fetal DNA to determine if the baby has certain types of genetic conditions such as Down syndrome. Such tests require about a tablespoon of blood.

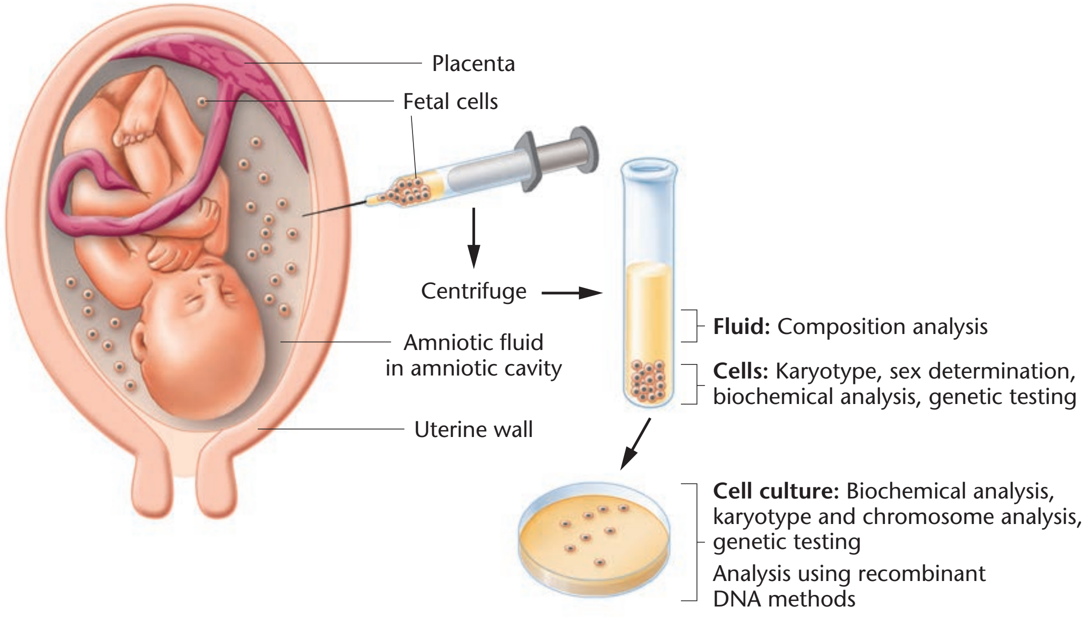

Placenta
Fetal cells
Centrifuge
Amniotic fluid in amniotic cavity
Uterine wall
Fluid: Composition analysis
Cells: Karyotype, sex determination, biochemical analysis, genetic testing
Cell culture: Biochemical analysis, karyotype and chromosome analysis, genetic testing
Analysis using recombinant DNA methods

FIGURE 22.5 For amniocentesis, the position of the fetus is first determined by ultrasound, and then a needle is inserted through the abdominal and uterine walls to recover amniotic fluid containing fetal cells for genetic or biochemical analysis.

DNA in the blood is sequenced to analyze **haplotypes**—contiguous segments of DNA that do not undergo recombination during gamete formation—that distinguish which cfDNA segments are maternal and which are from the fetus (see Figure 22.6). If a fetal haplotype contained a specific mutation, this would also be revealed by sequence analysis. Nearly complete fetal genome sequences have been assembled from maternal blood. These are developed by sequencing cfDNA fragments from maternal blood and comparing those fragments to sequenced genomes from the mother and father. Bioinformatics software is then used to organize the genetic sequences from the fetus in an effort to assemble the fetal genome. Currently, this technology does not capture the entire fetal genome; it results in an assembled genome sequence with segments missing. It has been shown, however, that **whole-genome sequencing (WGS)** (introduced in Chapter 21) of maternal plasma cfDNA can be used to accurately sequence the entire exome of a fetus.

Tests for fetal genetic analysis based on maternal blood samples started to arrive on the market in 2011. Sequenom of San Diego, California, was one of the first companies to launch such a test—*MaterniT® 21 PLUS*, a Down syndrome test that can also be used to test for trisomy 13 (Patau syndrome) and trisomy 18 (Edwards syndrome). MaterniT 21 PLUS analyzes 36-bp fragments of DNA to identify chromosome 21 from the fetus. Sequenom claims that this test is highly accurate with a false positive rate of just 0.2 percent. The test can be done as early as week 10 (about the same time CVS sampling can be performed, which is about 4 to 6 weeks earlier than amniocentesis can be performed). Several companies have followed the Sequenom approach. Nationwide, it has been estimated that the future market for these tests could be greater than $1 billion. As discussed later in this chapter, there are many ethical issues associated with prenatal genetic testing. Most insurance companies are not yet paying for WGS of maternal blood, which can cost as much as $2000 for a single test. Recently, California agreed to subsidize noninvasive prenatal testing for women through the state's genetic diseases program, which screens ~400,000 women each year.

Originally these noninvasive tests were offered to women older than 35 years of age, or if they were identified as at risk based on family history of birth-related complications. Now these tests are being marketed to women with low-risk pregnancies as well, and their value, given the cost, has been questioned. Recent figures indicate that sales for these tests exceeded $600 million annually, and this number is estimated to increase fourfold by 2022.

In Section 22.9 we will discuss preconception testing and recent patents for computing technologies designed to predict the genetic potential of offspring (destiny tests).

Genetic Testing Using Allele-Specific Oligonucleotides

A classic method of genetic testing was **restriction fragment length polymorphism (RFLP) analysis.** Historically RFLP analysis was the primary method to detect **sickle-cell anemia.** As discussed previously (Chapter 14), this disease is an autosomal recessive condition common in people with family origins in areas of West Africa, the Mediterranean basin, parts of the Middle East, and India. It is caused by a single amino acid substitution in the β-globin protein, as a consequence of a single-nucleotide substitution

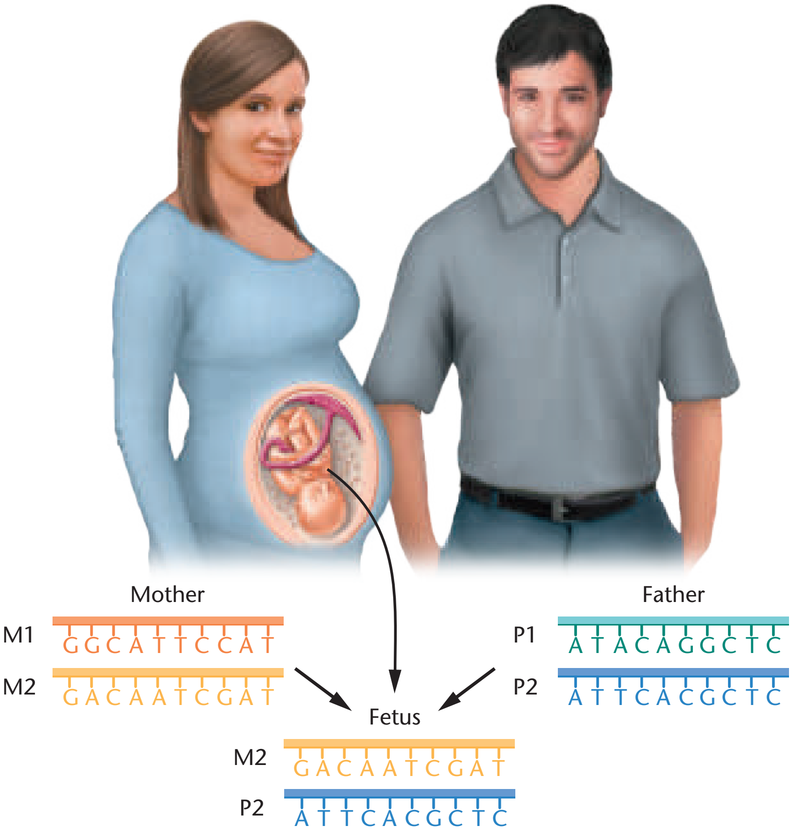


FIGURE 22.6 Deducing fetal genome sequences from maternal blood. For any given chromosome, a fetus inherits one copy of a haplotype from the mother (maternal copies, M1 or M2) and another from the father (paternal copies, P1 or P2). For simplicity, a single-stranded sequence of DNA from each haplotype is shown. These haplotype sequences can be detected by WGS. Here the fetus inherited haplotypes M2 and P2 from the mother and father, respectively. DNA from the blood of a pregnant woman would contain paternal haplotypes inherited by the fetus (P2, blue), maternal haplotypes that are not passed to the fetus (M1, orange), and maternal haplotypes that are inherited by the fetus (M2, yellow). The maternal haplotype inherited by the fetus (M2) would be present in excess amounts relative to the maternal haplotype that is not inherited (M1).

in the corresponding gene. The single-nucleotide substitution also eliminates a restriction site in the β-globin gene for the restriction enzymes *Mst* II and *Cvn* I. As a result, the mutation alters the pattern of restriction fragments seen on Southern blots. These differences in restriction sites could be used to diagnose sickle-cell anemia prenatally and to establish the parental genotypes and the genotypes of other family members who may be heterozygous carriers of this condition.

But only about 5 to 10 percent of all point mutations can be detected by RFLP analysis because most mutations occur in regions of the genome that do not contain restriction enzyme sites. However, since the Human Genome Project (HGP) was completed and many more disease-associated mutations became known, geneticists can now employ PCR and synthetic oligonucleotides to detect these mutations, including the use of synthetic DNA probes known as **allele-specific oligonucleotides (ASOs).** This rapid, inexpensive, and accurate technique is used to diagnose a wide range of genetic disorders caused by point mutations. In contrast to RFLP analysis, which is limited to cases for which a mutation changes a restriction site, ASOs detect single-nucleotide changes **(single-nucleotide polymorphisms** or **SNPs).**

An ASO is a short, single-stranded fragment of DNA designed to hybridize to a complementary specific allele in the genome. Under proper conditions, an ASO will hybridize only with its complementary DNA sequence and not with other sequences, even those that vary by as little as a single nucleotide.

Genetic testing using ASOs and PCR analysis is available to screen for many disorders, including sickle-cell anemia. In the case of sickle-cell anemia screening, DNA is extracted (either from a maternal blood sample or from fetal cells obtained by amniocentesis), and a region of the β-globin gene is amplified by PCR. A small amount of the amplified DNA is spotted onto strips of a DNA-binding membrane, and each strip is hybridized to an ASO synthesized to resemble the relevant sequence from either a normal or mutant β-globin gene. The ASO is tagged with a molecule that is either radioactive or fluorescent, to allow for visualization of the ASO hybridized to DNA on the membrane. Figure 22.7 illustrates the principle behind this approach. This

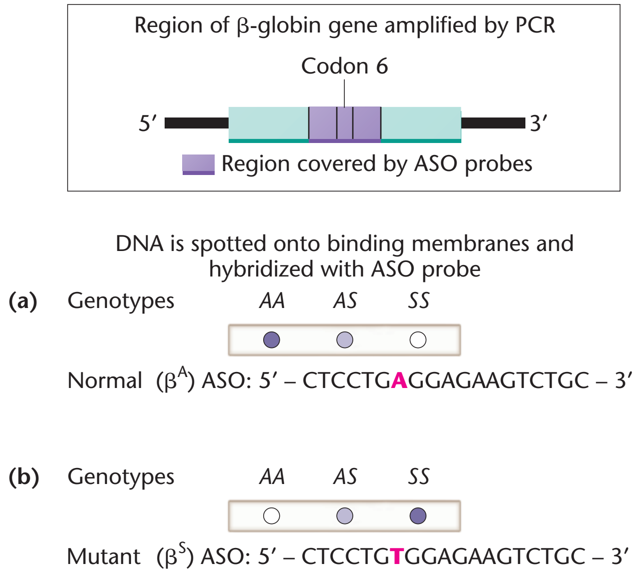

Region of β-globin gene amplified by PCR
Codon 6
5′
3′
Region covered by ASO probes
DNA is spotted onto binding membranes and hybridized with ASO probe
(a) Genotypes AA AS SS
Normal (β^A) ASO: 5′ – CTCCTGAGGAGAAGTCTGC – 3′
(b) Genotypes AA AS SS
Mutant (β^S) ASO: 5′ – CTCCTGTGGAGAAGTCTGC – 3′

FIGURE 22.7 Allele-specific oligonucleotide (ASO) testing for sickle-cell anemia. (a) Results observed if the three possible β-globin genotypes are hybridized to an ASO for the normal β-globin allele: *AA*-homozygous individuals have normal hemoglobin that has two copies of the normal β-globin gene and will show heavy hybridization; *AS*-heterozygous individuals carry one normal β-globin allele and one mutant allele and will show weaker hybridization; *SS*-homozygous sickle-cell individuals carry no normal copy of the β-globin gene and will show no hybridization to the ASO probe for the normal β-globin allele. (b) Results observed if DNA for the three genotypes are hybridized to the probe for the sickle-cell β-globin allele: no hybridization by the *AA* genotype, weak hybridization by the heterozygote (*AS*), and strong hybridization by the homozygous sickle-cell genotype (*SS*).

rapid, inexpensive, and accurate technique is used to diagnose a wide range of genetic disorders caused by point mutations.

Although ASO testing is highly effective, SNPs can affect ASO probe binding leading to false positive or false negative results that may not reflect a genetic disorder, particularly if precise hybridization conditions are not used. Sometimes DNA sequencing is carried out on amplified gene segments to confirm identification of a mutation.

Because ASO testing makes use of PCR, only small amounts of DNA are required for analysis. As a result, ASO testing is ideal for **preimplantation genetic diagnosis (PGD).** PGD involves the genetic analysis of single cells from embryos created by *in vitro* fertilization (IVF) (**Figure 22.8**). PGD has been used for over 25 years, typically when there is concern about a particular genetic defect.

When sperm and eggs are mixed to create zygotes, the early-stage embryos are grown in culture. A single cell can be removed from an early-stage embryo using a vacuum pipette to gently aspirate one cell away from the embryo (Figure 22.8, top). This could possibly kill the embryo, but if it is done correctly, the embryo will often continue to divide normally. DNA from the single cell is then typically analyzed by FISH (for chromosome analysis) or by ASO testing (Figure 22.8, bottom). The genotypes for each early-stage embryo can be tested to decide which embryos will be implanted into the uterus.

Any alleles that can be detected by ASO testing can be used during PGD. Sickle-cell anemia, cystic fibrosis, and dwarfism are often tested for by PGD, but alleles for many other conditions are also often analyzed. In theory, PGD should improve embryo implantation success rates and reduce miscarriages for couples—and success rates have improved, particularly for older women undergoing IVF. But this turns out not to be true for all couples because PGD cannot be used for identifying epigenetic changes that affect fertility. As we learn more about epigenetic influences on fertilization, it is expected that PGD will be expanded to incorporate epigenetic analysis in the future.

Also, as you will learn in Sections 22.5 and 22.6, it is now possible to carry out whole-genome sequencing (WGS) on individual cells. This method is now being applied for PGD of single cells from an embryo created by IVF.

NOW SOLVE THIS

22.2 The DNA sequence surrounding the site of the sickle-cell mutation in the β-globin gene, for normal and mutant genes, is as follows:

Normal DNA

5′-GACTCCTGAGGAGAAGT-3′
3′-CTGAGGACTCCTCTTCA-5′

Sickle-cell DNA

5′-GACTCCTGTGGAGAAGT-3′
3′-CTGAGGACACCTCTTCA-5′

Each type of DNA is denatured into single strands and applied to a DNA-binding membrane. The membrane containing the two spots is hybridized to an ASO of the sequence

5′-GACTCCTGAGGAGAAGT-3′

Which spot, if either, will hybridize to this probe?

- **HINT:** *This problem asks you to analyze results of an ASO test. The key to its solution is to understand that ASO analysis is done under conditions that allow only identical nucleotide sequences to hybridize to the ASO on the membrane.*

For more practice see Problem 8.

Genetic Testing Using Microarrays

ASO analysis is an effective method of screening for one, or a small number, of mutations within a gene. However, there is a significant demand for genetic tests that detect complex mutation patterns or previously unknown mutations in a single gene associated with genetic diseases and cancers. For example, the gene that is responsible for cystic fibrosis (the *CFTR* gene) contains 27 exons and encompasses 250 kilobases of genomic DNA. Of the 1000 known mutations of the *CFTR* gene, about half are point mutations, insertions,

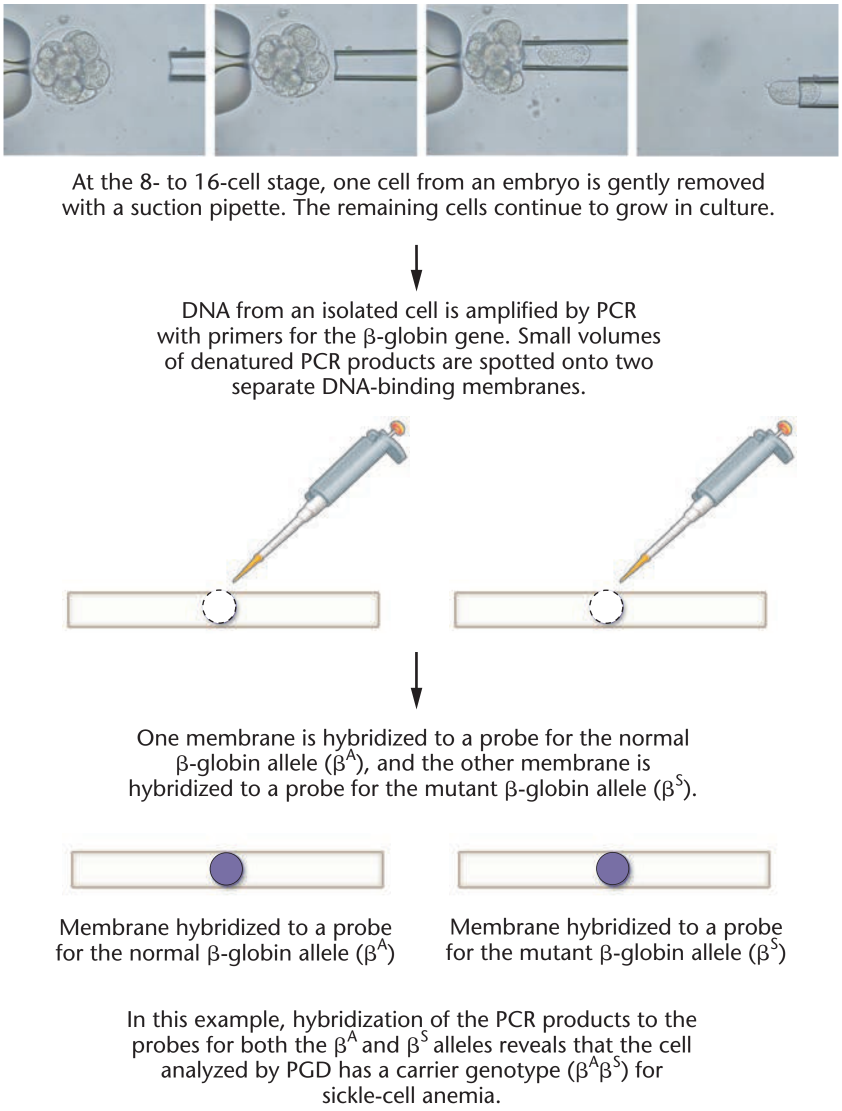

At the 8- to 16-cell stage, one cell from an embryo is gently removed with a suction pipette. The remaining cells continue to grow in culture.
DNA from an isolated cell is amplified by PCR with primers for the β-globin gene. Small volumes of denatured PCR products are spotted onto two separate DNA-binding membranes.
One membrane is hybridized to a probe for the normal β-globin allele (β^A), and the other membrane is hybridized to a probe for the mutant β-globin allele (β^S).
Membrane hybridized to a probe for the normal β-globin allele (β^A)
Membrane hybridized to a probe for the mutant β-globin allele (β^S)
In this example, hybridization of the PCR products to the probes for both the β^A and β^S alleles reveals that the cell analyzed by PGD has a carrier genotype ($\beta^A\beta^S$) for sickle-cell anemia.

FIGURE 22.8 A single cell from an early-stage human embryo created by *in vitro* fertilization can be removed and subjected to preimplantation genetic diagnosis (PGD) by ASO testing. DNA from the cell is isolated, amplified by PCR with primers specific for the gene of interest, and then subjected to ASO analysis as shown in Figure 22.7. In this example, a region of the β-globin gene was amplified and analyzed by ASO testing to determine the sickle-cell genotype for this cell.

and deletions—and they are widely distributed throughout the gene. Similarly, over 500 different mutations are known to occur within the tumor suppressor *p53* gene, and any of these mutations may be associated with, or predispose a patient to, a variety of cancers. In order to screen for mutations in these genes, comprehensive, high-throughput methods are required.

From earlier in the text (see Chapter 21), recall that one high-throughput screening technique is based on the use of **DNA microarrays** (also called DNA or gene chips). The numbers and types of DNA sequences on a microarray are dictated by the type of analysis that is required. For example, each spot or field (sometimes also called a feature) on a microarray might contain a DNA sequence derived from each member of a gene family, sequence variants from one or several genes of interest, or a sequence derived from each gene in an organism's genome.

In the recent past, DNA microarrays have been used for a wide range of applications, including the detection of mutations in genomic DNA and the detection of gene-expression patterns in diseased tissues. However, in the near future, whole-genome sequencing, exome sequencing, and RNA sequencing are expected to replace most applications involving microarrays and render this technology obsolete.

But because of the impact microarrays have had on genetic testing, it is still valuable to discuss applications of this method. What makes DNA microarrays so useful is the immense amount of information that can be simultaneously generated from a single array. DNA microarrays the size of postage stamps (just over 1 cm square) can contain up to 500,000 different fields, each representing a different DNA sequence. Human genome microarrays containing probes for most human genes are available, including many disease-related genes, such as the *p53* gene, which is mutated in a majority of human cancers, and the *BRCA1* gene, which, when mutated, predisposes women to breast cancer and men to breast and prostate cancer.

In addition to testing for mutations in single genes, DNA microarrays can include probes that detect SNPs. SNPs crop up in an estimated 15 million positions in the genome where these single-based changes reveal differences from one person to the next. SNP sequences as probes on a DNA microarray allow scientists to simultaneously screen thousands of mutations that might be involved in single-gene diseases as well as those involved in disorders exhibiting multifactorial inheritance. This technique, known as **genome scanning,** makes it possible to analyze a person's DNA for dozens or even hundreds of disease alleles, including those that might predispose the person to heart attacks, asthma, diabetes, Alzheimer disease, and other genetically defined disease subtypes. Genome scans are now occasionally used when physicians encounter patients with chronic illnesses where the underlying cause cannot be diagnosed.

In contrast to genome scanning microarrays that detect mutations in DNA, **gene-expression microarrays** detect gene-expression patterns for specific genes. This can be an effective approach for diagnosing genetic diseases because the progression of a tissue from a healthy to a diseased state is almost always accompanied by changes in mRNA expression of hundreds to thousands of genes. Gene-expression microarrays may contain probes for only a few specific genes thought to be expressed differently in different cell types or may contain probes representing each gene in the genome. Although microarray techniques provide novel information about gene expression, it should be emphasized that DNA microarrays do not directly provide us with information about protein levels in a cell or tissue. We often infer what predicted protein levels may be based on mRNA expression patterns, but this may not always be accurate.

In one type of gene-expression microarray analysis, mRNA is isolated from two different cell or tissue types—for example, normal cells and cancer cells arising from the same cell type (Figure 22.9). The mRNA samples contain transcripts from each gene that is expressed in that cell type. Some genes are expressed at higher levels than others. The expression level of each mRNA can be used to develop a *gene-expression profile* that is characteristic of the cell type. To do this, isolated mRNA molecules are converted into cDNA molecules, using reverse transcriptase. The cDNAs

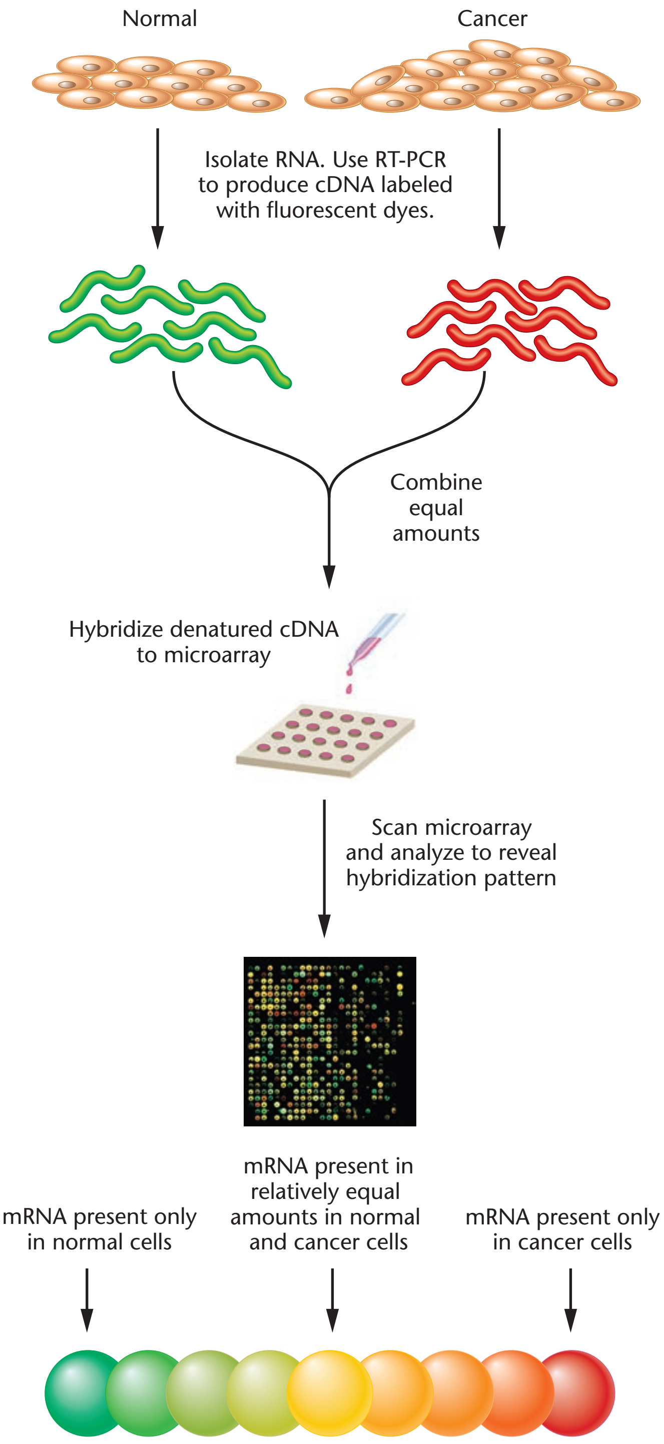

Normal
Cancer
Isolate RNA. Use RT-PCR to produce cDNA labeled with fluorescent dyes.
Combine equal amounts
Hybridize denatured cDNA to microarray
Scan microarray and analyze to reveal hybridization pattern
mRNA present only in normal cells
mRNA present in relatively equal amounts in normal and cancer cells
mRNA present only in cancer cells

FIGURE 22.9 Microarray procedure for analyzing gene expression in normal and cancer cells. The method shown here is based on a two-channel microarray in which cDNA samples from the two different tissues are competing for binding to the same probe sets. The colors of the dots on an expression microarray represent levels of gene expression in a "heat map." In the heat map shown here, green dots represent genes expressed only in one cell type (i.e., the normal cells), and red dots represent genes expressed only in another cell type (i.e., the cancer cells). Intermediate colors represent different levels of expression of the same gene in the two cell types. (Only a small portion of the complete DNA microarray is shown.)

from the normal cells are tagged with fluorescent dye-labeled nucleotides (for example, green), and the cDNAs from the cancer cells are tagged with a different fluorescent dye-labeled nucleotide (for example, red).

The labeled cDNAs are mixed together and applied to a DNA microarray. The cDNA molecules bind to complementary single-stranded probes on the microarray but not to other probes. Keep in mind that each field or feature does not consist of just one probe molecule but rather contains thousands of copies of the probe. After washing off the nonbinding cDNAs, scientists scan the microarray with a laser, and a computer captures the fluorescent image pattern for analysis. The pattern of hybridization appears as a series of colored dots, with each dot corresponding to one field of the microarray.

This color pattern representation of results is often referred to as a *heat map*, because the color (or intensity of brightness) of a particular spot provides a sensitive measure of the relative levels of each cDNA in the mixture. In the example shown in Figure 22.9, if an mRNA is present only in normal cells, the probe representing the gene encoding that mRNA will appear as a green dot because only "green" cDNAs have hybridized to it. Similarly, if an mRNA is present only in the cancer cells, the microarray probe for that gene will appear as a red dot. If both samples contain the same cDNA, in the same relative amounts, both cDNAs will hybridize to the same field, which will appear yellow. Intermediate colors indicate that the cDNAs are present at different levels in the two samples.

Gene-expression microarray analysis has revealed that certain cancers have distinct patterns of gene expression and that these patterns correlate with factors such as the cancer's stage, clinical course, or response to treatment. For example, scientists examined gene expression in both normal white blood cells and in cells from a white blood cell cancer known as *diffuse large B-cell lymphoma (DLBCL)*. About 40 percent of patients with DLBCL respond well to chemotherapy and have long survival times. The other 60 percent respond poorly to therapy and have short survival.

The investigators assayed the expression profiles of 18,000 genes and discovered that there were two types of DLBCL, with almost inverse patterns of gene expression (**Figure 22.10**). One type of DLBCL, called GC B-like, had an expression pattern dramatically different from that of a second type, called activated B-like. Patients with the activated B-like pattern of gene expression had much lower survival rates than patients with the GC B-like pattern. The researchers concluded that DLBCL is actually two different diseases with different outcomes.

Once this type of profiling analysis is introduced into routine clinical use, it may be possible to adjust therapies for each group of cancer patients and to identify new specific treatments based on gene-expression profiles. Similar gene-expression profiles have been generated for many other cancers, including breast, prostate, ovarian, and colon cancer, providing tremendous insight into both substantial and subtle variations in genetic diseases.

Several companies are now promoting **nutrigenomics** services in which they claim to use gene-expression analysis to identify allele polymorphisms and gene-expression patterns for genes involved in nutrient metabolism. For

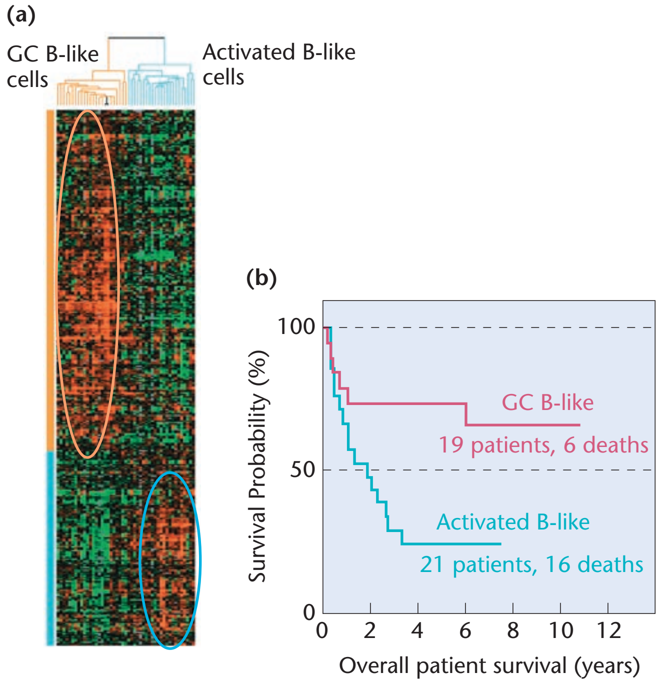


FIGURE 22.10 (a) Gene-expression microarray results analyzing 18,000 genes expressed in DLBCL lymphocytes. Each row represents a summary of the gene expression from one particular gene; each column represents data from one cancer patient's sample. In this heat map, the colors represent ratios of relative gene expression compared to normal control cells. Red represents expression greater than the mean level in controls, green represents expression lower than in the controls, and the intensity of the color represents the magnitude of difference from the mean. In this summary analysis, the cancer patients' samples are grouped by how closely their gene-expression profiles resemble each other. The cluster of cancer patients' samples marked with an orange oval at the top of the figure are GC B-like DLBCL cells. The blue oval cluster contains samples from cancer patients within the activated B-like DLBCL group. (b) Gene-expression profiling and survival probability. Patients with activated B-like profiles have a poorer chance of survival (16 in 21) than those with GC B-like profiles (6 in 19). Data such as these demonstrate the value of microarray analysis for diagnosing disease conditions.

example, polymorphisms in genes such as that for apolipoprotein A (*APOA1*), involved in lipid metabolism, and for methylenetetrahydrofolate reductase (*MTHFR*), involved in metabolism of folic acid, have been implicated in cardiovascular disease. Nutrigenomics companies claim that analysis of a patient's DNA for genes such as these enables them to judge whether allele variations or gene-expression profiles warrant dietary changes to potentially improve that person's health and reduce the risk of diet-related diseases.

Applications of Gene-Expression Microarrays and Next-Generation Sequencing for Pathogen Identification

Among their many applications, microarrays have provided infectious disease researchers with powerful tools for studying pathogens. Genotyping microarrays have been used to identify strains of emergent viruses, such as the virus that causes the highly contagious condition called severe acute respiratory syndrome (SARS) as well as the H5N1 avian influenza virus, the cause of bird flu, which has killed people in Asia, leading to the slaughter of over 80 million chickens and causing concern about possible pandemic outbreaks.

Researchers are using whole-genome transcriptome analysis (see Chapter 21) to investigate genes that are important for pathogen infection and replication. In this approach, host cells are infected *in vitro* with bacteria, yeast, protist, or viral pathogens, and then gene-expression microarrays are used to analyze pathogen gene-expression profiles. Patterns of pathogen gene activity during infection and replication are useful for identifying pathogens and understanding mechanisms of infection. This strategy primarily informs researchers about how a pathogen responds to its host. But, of course, a primary goal of infectious disease research is to prevent infection. Gene-expression profiling can identify important pathogen genes and the proteins they encode. The latter may prove to be useful targets for subunit vaccine development or for drug treatment strategies to prevent or control infectious disease.

Similarly, researchers are evaluating host responses to pathogens. The application of this approach has been accelerated in part by the need to develop pathogen-detection strategies for military and civilian use—both for detecting outbreaks of naturally emerging pathogens such as SARS and avian influenza and for potential detection of outbreaks such as anthrax (caused by the bacterium *Bacillus anthracis*) that could be the result of a bioterrorism event. Host-response gene-expression profiles are developed by exposing a host to a pathogen and then using gene-expression microarrays to analyze host gene-expression patterns.

Figure 22.11 shows the different gene-expression profiles for mice following exposure to *Neisseria meningitidis*, the SARS virus, or *E. coli*. In this example, several

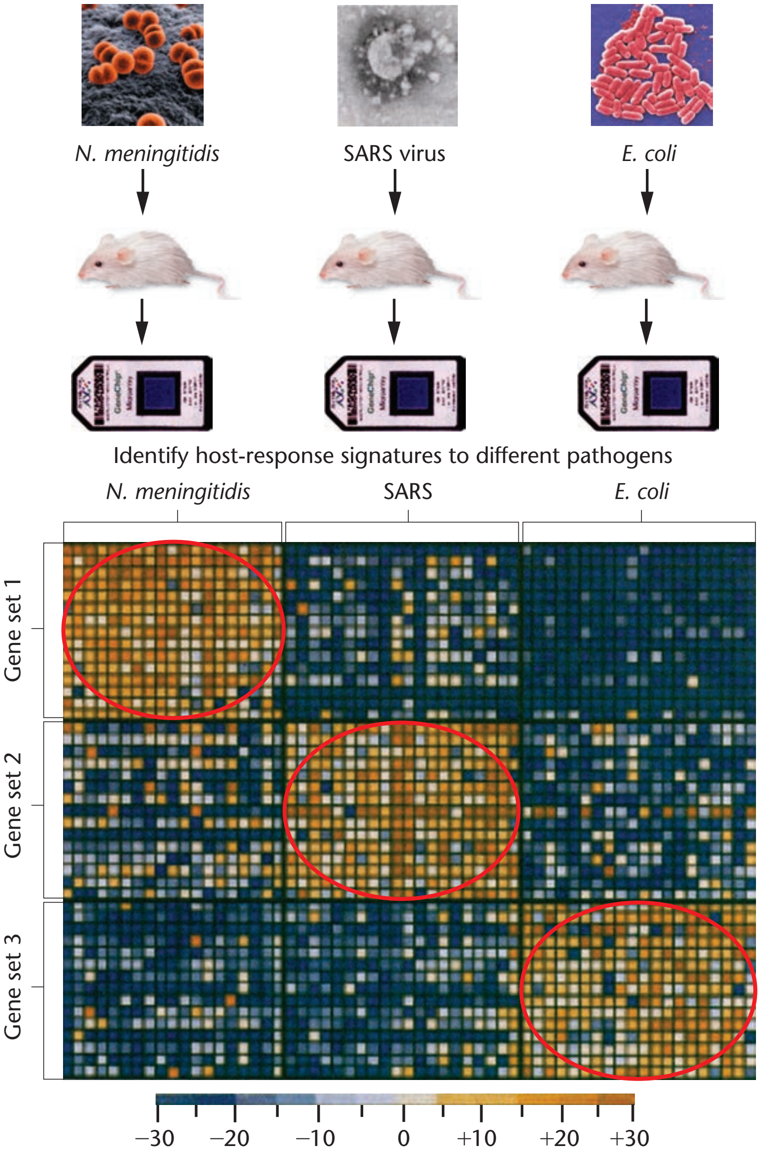


FIGURE 22.11 Gene-expression microarrays can reveal host-response signatures for pathogen identification. In this example, mice were infected with different pathogens: *Neisseria meningitidis*, the virus that causes severe acute respiratory syndrome (SARS), and *E. coli*. Mouse tissues were then used as the source of mRNA for gene-expression microarray analysis. Increased expression compared to uninfected control mice is shown in shades of yellow. Decreased expression compared to uninfected controls is indicated in shades of blue. Notice that each pathogen elicits a somewhat different response in terms of which major clusters of host genes are activated by pathogen infection (red circles).

genes are upregulated or downregulated by each pathogen. Notice, however, how each pathogen strongly induces different prominent clusters of genes that reveal a host gene-expression response to the pathogen and provide a signature of pathogen infection. Comparing such host gene-expression profiles following exposure to different pathogens provides researchers with a way to quickly diagnose and classify infectious diseases. Scientists are developing databases of both pathogen and host-response expression profile data that can be used to identify pathogens efficiently.

Increasingly, next-generation sequencing (NGS) and third-generation sequencing (TGS) methods are being used for pathogen identification. Whole-genome sequencing (WGS) is a quick way to identify pathogenic bacteria, viruses, or other microorganisms, and it is also very valuable for tracking genetic variations in microorganisms. This allows public health officials to track the evolution of changes in microbes to help develop proactive approaches to combat outbreaks of new, deadly strains. NGS provided a large body of data on evolving strains of Ebola virus during the 2014—15 outbreak in West Africa, information that was essential for advancing vaccine development.

From the first confirmed hospitalized case of a woman with Ebola in Sierra Leone in late May 2014, blood samples from patients testing positive for Ebola were shipped to a lab in Cambridge, Massachusetts, for sequencing. By mid-June 149 blood samples from 78 patients had been processed for deep sequencing. Alignment of these sequences yielded 78 confirmed cases of Ebola and provided 99 full-length genomes. New mutations were mapped to a reference genome. From genomic analysis, it was estimated that strains from the current outbreak diverged from a Central African version of the Ebola virus less than 10 years ago.

Avian influenza, Middle East respiratory syndrome (MERS), and Zika virus are among other pathogens that have been identified and characterized by WGS.

Screening the Genome for Genes or Mutations You Want

While we have thus far focused on genetic testing and identifying genes involved in disease, genomic analysis is also revealing genome diversity that confers beneficial attributes or phenotypes to humans. This can have a role in medical diagnosis because it allows scientists and physicians to understand why genetic differences account for resistance to certain diseases in some individuals compared to others.

For instance, researchers at the Broad Institute in Boston were studying elderly, overweight individuals who by all conventional medical diagnostic approaches should have shown symptoms of diabetes, yet they were not diabetic. Instead of seeking diabetes-causing mutations, the Broad group employed genomic analysis to search for mutations associated with protection from diabetes. Their efforts were rewarded when they determined that individuals with loss-of-function mutations in the *SLC30A8* gene (*solute carrier family 30, member 8* gene, which encodes a zinc transport protein involved in insulin secretion) are 65 percent less likely to develop diabetes even when they have highly associated risk factors such as obesity.

In this spirit, increasingly geneticists are analyzing “natural or healthy knockouts”—the fortunate few individuals who may lack a specific gene or have a mutation in a disease-causing gene that provides a health benefit, such as protecting against development of a particular disease. Identifying such genetic variations may make it easier to help combat infections and disease.

For example, through mutation many viruses that infect humans can evade drugs used to combat them. But these same viruses are defenseless against a rare mutation in the human gene *ISG15*. Individuals with mutations in *ISG15* fight off many if not most viruses. (Estimates suggest that less than 1 person in 10 million has this mutation.) *ISG15* mutations knock out a function that helps to dampen inflammation, so individuals with this mutation have a heighted inflammatory system, which helps fight off viruses. It is thought that this elevated response prevents viruses from replicating to levels that typically cause illness.

Based on this knowledge, researchers are trying to develop drugs that might mimic the effects of the *ISG15* mutation as a future treatment strategy. As you will learn later in text (in Special Topic Chapter 5—Gene Therapy), this strategy has been used to mimic naturally occurring mutations in the *CCR5* receptor gene that provides a rare subset of individuals (~ 1 percent of northern Europeans) with complete immunity to HIV infection.

A project called the Exome Aggregation Consortium (ExAC) is cataloging genetic variation from exome sequences of more than 60,000 individuals from diverse ancestries with the purpose of identifying naturally occurring gene knockouts. People with such knockouts or those who carry disease-causing genes but don’t develop a particular illness are of significant interest to geneticists. Clearly this is an area of research that will continue to advance rapidly.

22.5 Genetic Analysis of Individual Genomes

Because of the relatively low cost of quickly and accurately sequencing individual genomes—what we call **personal genomics**—the ways that scientists and physicians evaluate a person’s genetic information is rapidly changing.

Whole-genome sequencing (WGS) is being utilized in medical clinics at an accelerating rate. Many major hospitals around the world are setting up clinical sequencing facilities for use in screening for the causes of rare diseases.

Recently, WGS has provided new insights into the genetics of anorexia, Alzheimer disease, and autism, among other disorders. Already there have been some very exciting success stories whereby WGS of individual genomes has led to improved treatment of diseases in children and adults. For example, native Newfoundlanders have one of the highest incidences in the world of *arrhythmogenic right ventricular dysplasia/cardiomyopathy* (*ARVD/C*), a rare condition in which affected individuals often have no symptoms but then die suddenly from irregular electrical impulses within the heart. Through individual genome sequencing, a mutation in the *AVRD5* gene has been identified as the cause of such cases of premature death. Of those with this mutation, approximately 50 percent of males and 5 percent of females die by age 40, and 80 percent of males and 20 percent of females die by age 50. Individuals carrying this mutation are now being implanted with internal cardiac defibrillators that can restart their hearts if electrical impulses stop or become irregular.

Diseases that are caused by multiple genes are much harder to diagnose and treat based on sequencing data. For example, WGS of individuals affected by **autism spectrum disorder (ASD)** has revealed the involvement of more than 100 different genes. The genetics of ASD is particularly complex because of the broad range of phenotypes associated with this disorder. While WGS of individuals affected by ASD has revealed inherited mutations, it has also identified sporadic *de novo* mutations. In the future, sequence-based knowledge of these mutations may help physicians develop patient-specific treatment strategies.

Recall our introduction of the concept of **whole-exome sequencing (WES)** (see Chapter 21). This alternative to WGS has also produced promising results in clinical settings. For example, from the time he was born, Nicholas Volmer had to live with unimaginable discomfort from an undiagnosed condition that was causing intestinal fistulas (holes from his gut to outside of his body) that were leaking body fluids and feces and requiring constant surgery. By 3 years of age, Nicholas had been to the operating room more than 100 times. A team at the Medical College of Wisconsin decided to have Nicholas's exome sequenced. Applying bioinformatics to compare his sequence to that of the general population, they identified a mutation in a gene on the X chromosome called *X-linked inhibitor of apoptosis* (*XIAP*). *XIAP* is known to be linked to another condition that can often be corrected by a bone marrow transplant. In 2010 a bone marrow transplant saved Nicholas's life and largely restored his health. Shortly thereafter the popular press described Nicholas as the first child saved by DNA sequencing.

Recently WGS and WES have been used to identify mutations of the *NGLY1* gene (which encodes a protein processing enzyme) that are associated with a very rare condition in children sharing certain development delays, liver disease, and a phenotype notable because of the inability to produce tears and thus the inability to cry. This list of confirmed patients with this mutation is less than two dozen, but this is the first time any definitive diagnosis has been available for these children despite a multitude of different clinical analysis and evaluations from doctors around the country.

The National Institutes of Health has an initiative called the *Undiagnosed Diseases Network*. Its goal is to use whole-exome and whole-genome sequencing to help diagnose rare and mysterious disease conditions of unknown genetic basis. In this program, exome sequences from an individual with a disorder can be compared to exome sequences from healthy family members and reference sequences to identify mutations that may be involved in the disease. To date, the program has diagnosed over 40 cases.

Unfortunately, even though a mutation responsible for a rare condition may have been identified, a cure or drug treatment has often not been developed for the disorder.

22.6 Genetic Analysis from Single Cells

We now have the ability to sequence the genome from a single cell! **Single-cell sequencing (SCS)** typically involves isolating genomic DNA from a single cell that is then subjected to *whole-genome amplification (WGA)* by PCR to produce sufficient DNA to be sequenced. Amplification of the genome to produce enough DNA for sequencing without introducing errors remains a major challenge that researchers are working on so that SCS can become a more reliable and accurate technique for genetic testing.

Genomic sequencing from single cells is valuable for analyzing both *somatic cell mutations* (for example, mutations that arise in somatic cells such as in a skin cancer, which are not heritable) as well as *germ-line mutations* (heritable mutations that are transmitted to offspring via gametes). Sequencing genomes from individual egg or sperm cells, especially for couples undergoing *in vitro* fertilization, can identify carrier conditions or specific germ-line mutations that could result in a genetic disorder in the offspring.

SCS allows scientists to explore genetic variations from cell to cell. These studies are revealing that different mutant genes can vary greatly between individual cells. In particular, cancer cells from a tumor often show genetic diversity, a fact that is increasingly being appreciated by researchers and clinicians. Understanding variations in genetic diversity and gene expression by individual cells within a tumor could lead to better and more specific treatment options.

The contribution of individual cells to the phenotype of a tissue or organ affected by a genetic disorder is increasingly of interest. **RNA sequencing (RNA-seq)** is becoming a powerful tool for transcriptome-wide analysis of genes expressed by cells within a population, thus allowing researchers to differentiate genetic variations between cells.

Until recently, researchers or clinicians had to analyze the genome and transcriptome from cells independently. But now, it is possible to isolate DNA and RNA from the same cells, sequence the DNA, and thanks to **single-cell RNA sequencing (scRNA-seq),** it is possible to sequence RNA from these same cells. This enables a comparison of the genes present in a cell and the relative levels of expression for each transcript encoded by the genome.

Many disease treatments are designed to target cells, such as those in a tumor, as if all cells are homogeneous in genotype and phenotype. In fact, often such cells are quite heterogeneous genetically. scRNA-seq is now being applied to reveal the heterogeneity of cell types in tumors and other conditions, to then help plan better treatment approaches based on genetics of the cell types and their relative abundance.

Sequencing DNA and RNA from the same cell type typically requires the use of PCR to amplify genomic DNA (to sequence DNA) or mRNA, which is reverse transcribed into cDNA and then subsequently incorporated into a library and sequenced. scRNA-seq also then provides a quantitative transcriptome analysis in which the relative levels of RNA expressed in a cell can be determined. scRNA-seq provides quantitative data about RNA expression, similar to gene-expression microarray analysis. But scRNA-seq reveals all transcripts expressed in a cell, whereas the transcripts analyzed by microarrays are limited by the probes present on the array. These are among the reasons why scRNA-seq is likely to eventually replace microarrays for transcriptome analysis.

As an example of scRNA-seq, **Figure 22.12** presents an analysis of innate lymphoid cells (ILCs). ILCs are a relatively recently identified group of immune cells that reside in bone marrow and other tissues of the body. ILCs can differentiate into a variety of different immune cell types. They resemble T lymphocytes (T cells), although they lack antigen-specific immune response capability, and play important roles in immunity and the regulation of inflammation. Abnormalities in ILCs are involved in conditions such as autoimmune diseases, allergic responses, and asthma. As a result, ILCs have emerged as important cellular targets for medical interventions designed to manipulate the immune system—approaches often called *immunotherapy*.

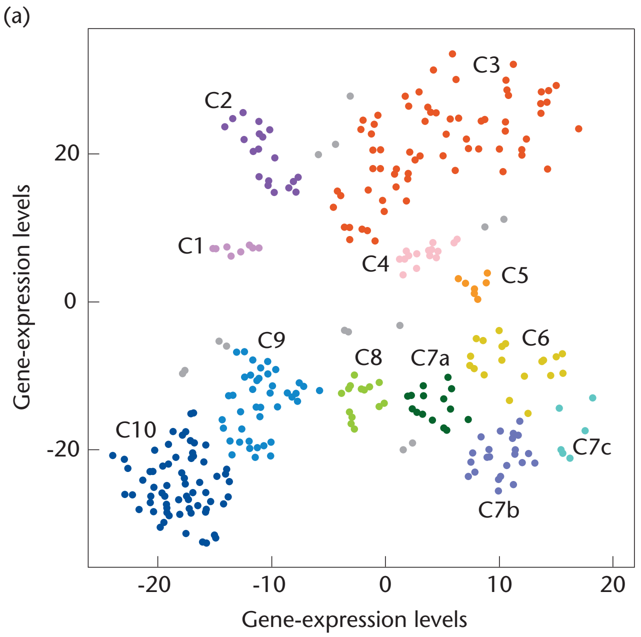
In the example shown in Figure 22.12, scRNA-seq of mouse bone marrow progenitor cells enabled identification of different subsets of ILCs by their RNA-expression patterns. In this experiment, levels of RNA expression are color-coded in “heat map” fashion similar to the way gene-expression results are displayed for microarray analysis.

Computational analysis applies algorithms that result in *clustering*—the grouping of cells based on similar patterns of gene-expression data. Such analysis reveals similarities in gene expression but significant differences in the transcriptomes of ILCs that, by phenotype, might appear to be the same. For example, in **Figure 22.12(b)** notice how the RNA-seq profiles of RNAs expressed by cells in cluster C10 are very different than the profiles of RNAs expressed by cells in cluster C3. This delineation of ILCs based on transcriptome analysis reveals important genetic differences and offers the potential to develop new therapeutic approaches to manipulate these cells and maximize their immune responses.

22.7 Genome-Wide Association Studies Identify Genome Variations That Contribute to Disease

Many of the genetic testing approaches we have discussed so far have focused on analyzing genes in individuals or relatively small numbers of people. Microarray-based genomic analysis and WGS have led geneticists to employ a powerful strategy called **genome-wide association studies (GWAS)** in their quest to analyze populations of people for disease genes.

GWAS of relatively large populations of people for diagnostic or prognostic purposes often enables scientists to identify multiple genes that may influence disease risk. During the past decade there has been a dramatic expansion in the number of GWAS being reported. For example, GWAS for autism, obesity, diabetes, macular degeneration, myocardial infarction, arthritis, hypertension, several cancers, bipolar disease, autoimmune diseases, Crohn disease, schizophrenia (a recent publication noted more than 100 genetic loci contributing to disease risk), amyotrophic lateral sclerosis, and multiple sclerosis are among the many GWAS that have been widely publicized in the scientific literature and popular press. Behavioral traits such as intelligence have also been analyzed by GWAS. For example, recently, a high-profile, controversial genome-wide association study reported genetic markers influencing cognitive ability and attempted to relate these markers to differences in educational attainment between people. Other studies have identified more than 50 genes that may influence intelligence.



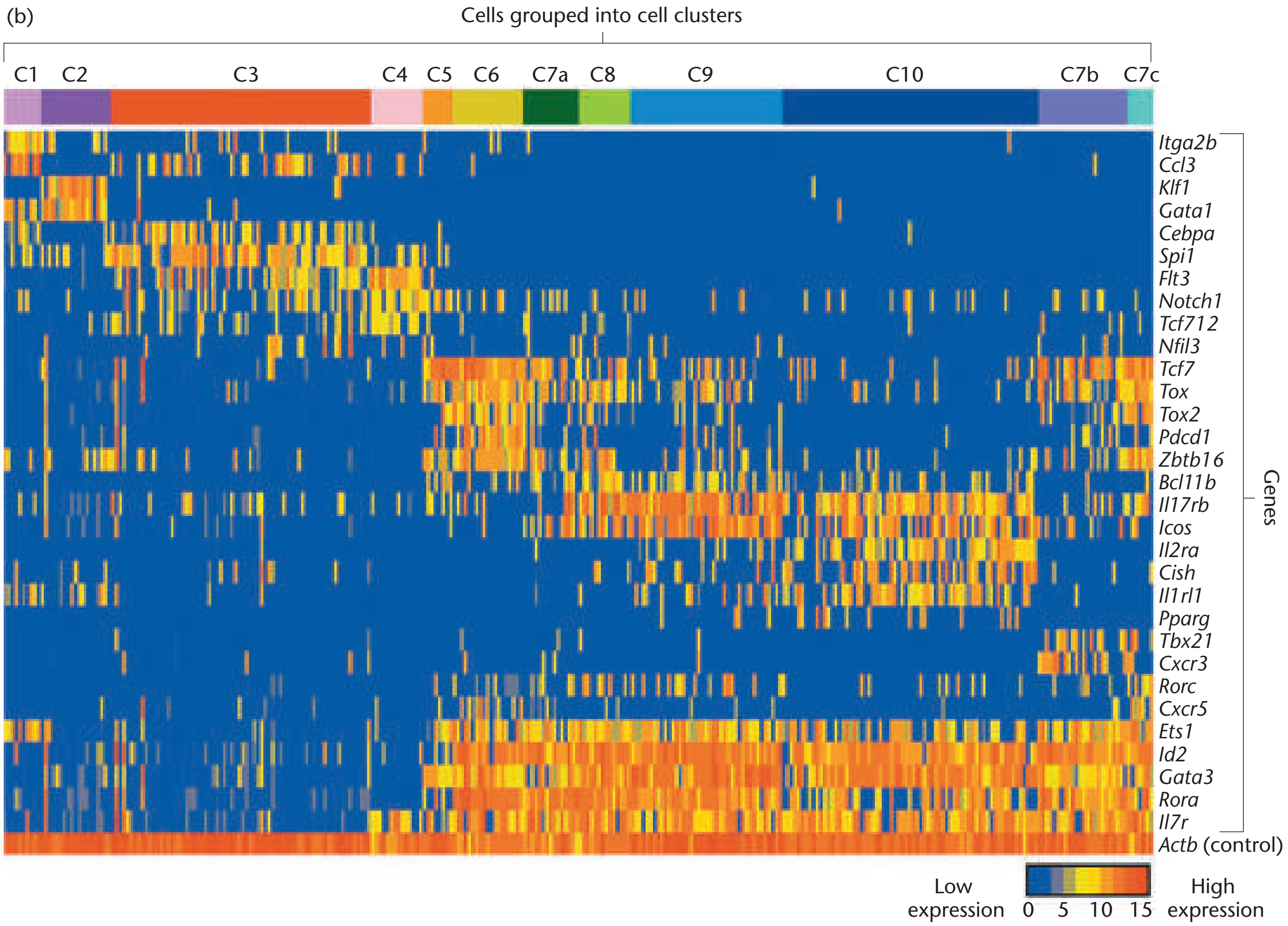


FIGURE 22.12 scRNA-seq analysis of innate lymphoid progenitor cells (ILCs). (a) A data plot of 325 ILCs from mouse bone marrow that were subjected to scRNA-seq. This data plot groups the ILCs into similar clusters based on shared gene-expression patterns. Each dot represents an individual cell, plotted against its expression levels for all genes analyzed by RNA-seq, and is color coded based on 10 genetically distinct clustering assignments (C1–C10). (b) A heat map displays RNA expression levels for selected genes in different cell clusters. Columns represent scRNA-seq data for each of the 325 different individual cells seen in part (a), grouped into the same 10 color-coded clusters. Each row displays a heat map for the expression levels of a specific gene (gene names shown to the far right) in all 325 cells. As a control, the bottom row of the heat map displays data for expression of β-actin mRNA; notice its relatively equal and high expression in all ILCs.

In a genome-wide association study, the genomes from several hundred, or several thousand (if available), unrelated individuals with a particular disease are analyzed, and results are compared with genomes of individuals without the disease. The goal is to identify genetic variations that may confer risk of developing the disease. Many GWAS involve large-scale use of SNP microarrays that can probe on the order of 500,000 SNPs to evaluate results from different individuals. Other approaches of GWAS use WGS to look for specific gene differences, evaluate CNVs, or search for changes in the epigenome, such as methylation patterns. By determining which SNPs, CNVs, or epigenome changes are present in individuals with the disease, scientists can calculate the disease risk associated with each variation. Analysis of GWAS results requires statistical analysis to predict the relative potential impact (association or risk) of a particular genetic variation on development of a disease phenotype.

Figure 22.13 shows a typical representation of one way that results from GWAS are commonly reported. The scatterplot representation, called a Manhattan plot, is used to display data with a large number of data points. Particular positions in the genome are plotted on the *x*-axis; in this case loci on each chromosome are plotted in a different color. The results of a genotypic association test are plotted on the *y*-axis. There are several ways that associations can be calculated. Shown here is a negative log of *p* values for loci determined to be significantly associated with a particular condition. The top line of this plot establishes a threshold value for significance. Marker sequences with significance levels exceeding the threshold *p* value of 10^{-5}, corresponding to 5.0 on the *y*-axis, are likely disease-related sequences.

One prominent study that brought the potential of GWAS to light involved research on 4587 patients in Iceland and the United States with a history of myocardial infarction (MI), and 12,767 control patients. This work was done with microarrays containing 305,953 SNPs. Among the most notable results, the study revealed variations in two tumor-suppressor genes (*CDKN2A* and *CDKN2B*) on

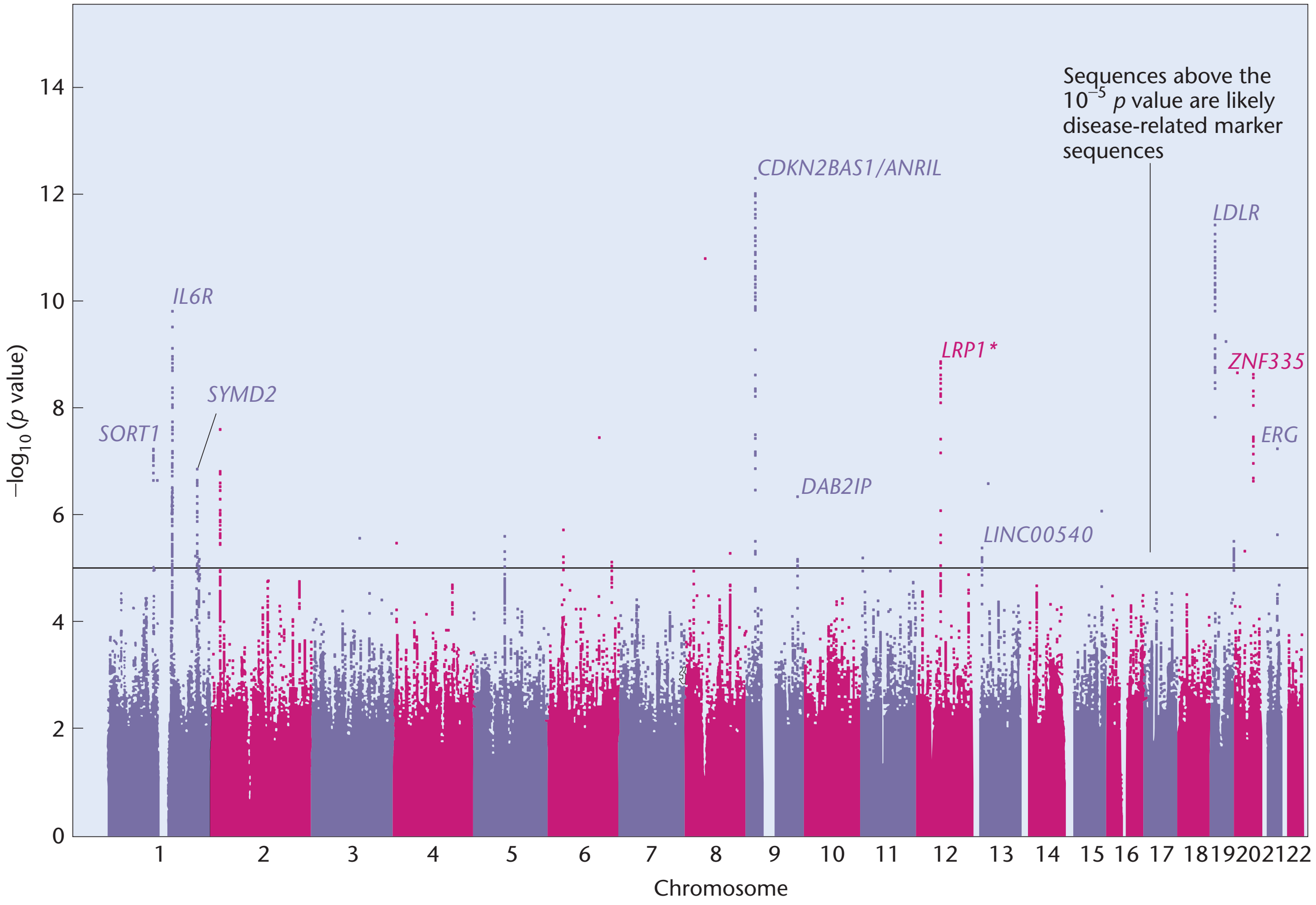


FIGURE 22.13 A genome-wide association study for Type 2 diabetes revealed 386,371 genetic markers, clustered here by chromosome number. Markers above the black line appeared to be significantly associated with the disease.

chromosome 9. Twenty-one percent of MI individuals were homozygous for deleterious mutations in both genes, and these individuals showed a 1.64 odds ratio of MI compared with noncarriers, including individuals homozygous for wild-type alleles. These variations were correlated with those in people of European descent, but interestingly, these same mutant alleles are not prominent in African-Americans. Does this mean that these genetic markers are not MI risk factors among the latter ethnicity?

Examples such as this raise questions and ethical concerns about patients' emotional responses to knowing about genetic risk data. For example, GWAS often reveal dozens of DNA variations, but many variations have only a modest effect on risk. How does one explain to a person that he or she has a gene variation that changes a risk difference for a particular disease from 12 to 16 percent over an individual's lifetime? What does this information mean? Similarly, if the sum total of GWAS for a particular condition reveals about 50 percent of the risk alleles, what are the other missing elements of heritability that may contribute to developing a complex disease? In some cases, risk data revealed by GWAS may help patients and physicians develop diet and exercise plans designed to minimize the potential for developing a particular disease.

GWAS are showing us that, unlike single-gene disorders, complex genetic disease conditions involve a multitude of genetic factors contributing to the total risk for developing a condition. We need such information to make meaningful progress in disease diagnosis and treatment, which is ultimately a major purpose of what GWAS are all about.

22.8 Synthetic Genomes and the Emergence of Synthetic Biology

Geneticists have long wondered about a fundamental question: "What is the minimum number of genes necessary (the *minimal genome*) to support life?" With the many advances in genomic analysis, the answer seems attainable. Today, scientists are even more interested in this question because the answer is expected to set the groundwork to create artificial cells or designer organisms based on genes encoded by a **synthetic genome** constructed in the laboratory. To do so we need a better understanding of the minimum number of genes, often referred to as the "core genes," required to support life.

The Minimal Genome: How Many Essential Genes Are Required by a Living Cell?

Scientists first considered the small genomes of obligate parasites to address this question. The bacterium *Mycoplasma genitalium*, a human parasitic pathogen, is among the simplest self-replicating cells known and thus has served as a model for understanding the minimal elements of a genome necessary for a self-replicating cell. *M. genitalium* can cause diseases in a wide range of hosts, including insects, plants, and humans. In humans, it causes genital infections.

In 1995, a team involving J. Craig Venter, Hamilton Smith, and Clyde Hutchison III sequenced the *M. genitalium* genome. At ~580 kb, with 525 genes, it is one of the smallest bacterial genomes known. In contrast, the 1.8-Mb genome of *Haemophilus influenzae* (the first bacterial genome sequenced) has 1815 genes. Which of the protein-encoding genes carried by *M. genitalium* constitute the minimal genome essential for life? Can we define life in terms of a minimal, essential, or "core" number of specific genes?

A combination of bioinformatics and experimental methods were used initially to answer these questions, including comparing the *M. genitalium* genome to the genomes of related species such as *M. pneumoniae* and *H. influenzae*. By comparing the nucleotide sequences of the *M. genitalium* genes with the *H. influenzae* genes, Venter and colleagues identified 256 genes whose sequence was similar enough to consider that they arose from a common ancestral gene; that is, they are orthologous. Thus, comparative genomics estimated that at least 256 genes might represent the minimum gene set essential for life.

But, could the number be more or less than 256? To answer this question, Venter and his colleagues applied an experimental approach. They used transposon-based methods to selectively mutate each gene in *M. genitalium* using the following rationale. Mutations in genes that produce a lethal phenotype indicate that the genes are essential, but nonessential mutated genes would not be lethal. They found that of the 525 genes in *M. genitalium*, about 375 genes were essential, thus constituting the minimum gene set in this bacterium.

Incidentally, there are symbiotic bacteria, such as a microbe called *Tremblaya princeps*, that contain only 120 protein-coding genes. However, *T. princeps* is not free-living, existing in a symbiotic relationship with an insect called a mealybug. Conceptually, the question of "how simple can a genome be?" continued to drive Venter and others to design experiments using synthetic genomic approaches to answer this question.

Design and Transplantation of a Synthetic Genome Defines the Minimal Bacterial Genome

In 2010, scientists from the J. Craig Venter Institute (JCVI) published the first report of a functional synthetic genome. In this approach they designed and chemically synthesized more than one thousand 1080-bp segments called cassettes covering the entire 1.08-Mb *Mycoplasma mycoides* genome [Figure 22.14(a)]. A homologous recombination technique was used to assemble the cassettes into

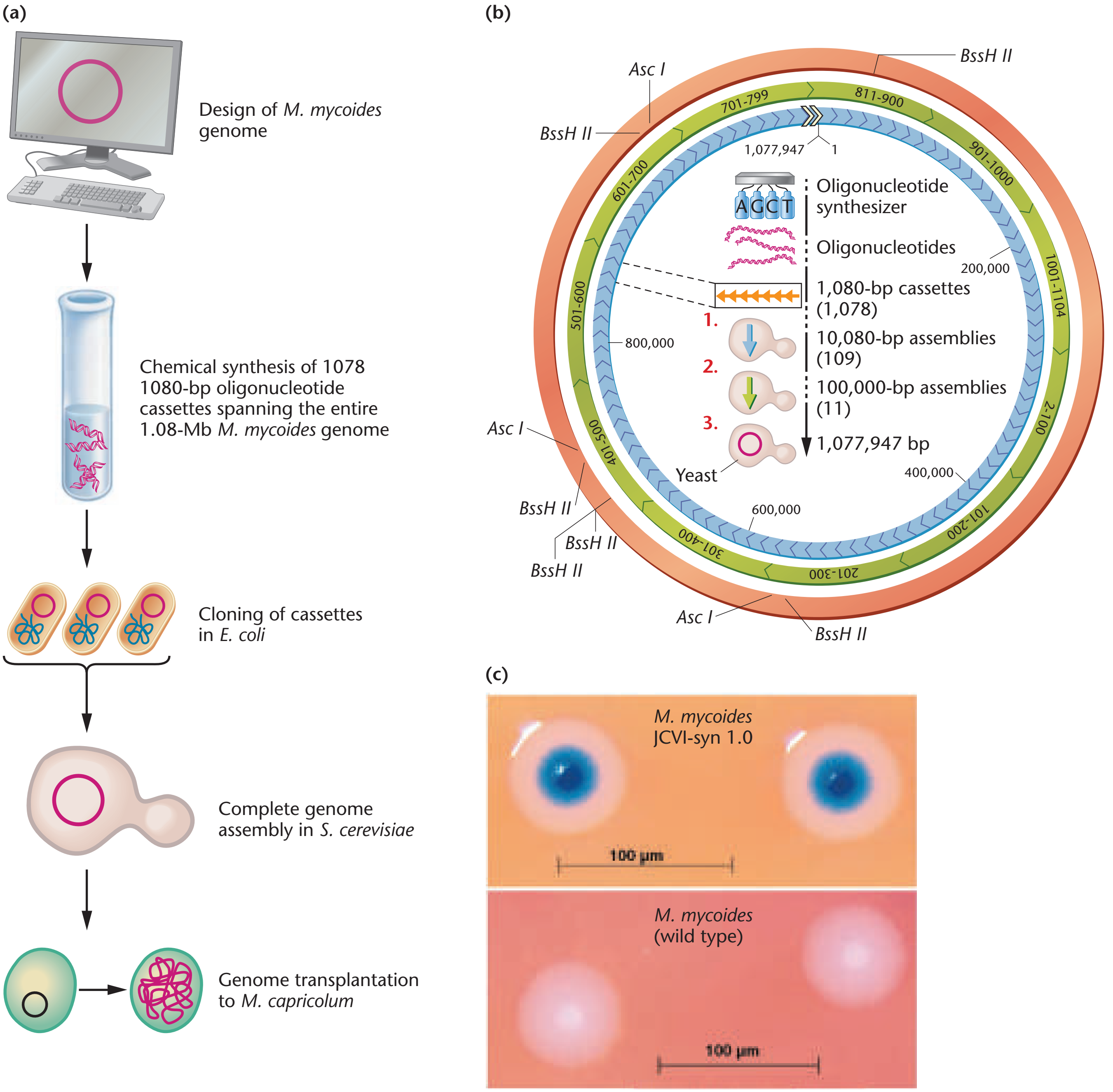

(a)
Design of *M. mycoides* genome
Chemical synthesis of 1078 1080-bp oligonucleotide cassettes spanning the entire 1.08-Mb *M. mycoides* genome
Cloning of cassettes in *E. coli*
Complete genome assembly in *S. cerevisiae*
Genome transplantation to *M. capricolum*
(b)
Asc I
BssH II
BssH II
701-799
811-900
901-1000
1001-1104
2-100
101-200
201-300
301-400
401-500
501-600
601-700
1,077,947
1
AGCT
Oligonucleotide synthesizer
Oligonucleotides
200,000
1,080-bp cassettes (1,078)
1.
10,080-bp assemblies (109)
2.
100,000-bp assemblies (11)
3.
1,077,947 bp
800,000
400,000
600,000
Yeast
Asc I
BssH II
BssH II
BssH II
Asc I
BssH II
(c)
M. mycoides JCVI-syn 1.0
100 µm
M. mycoides (wild type)
100 µm

FIGURE 22.14 Building a synthetic version of the 1.08-Mb *Mycoplasma mycoides* genome. (a) Overview of the approach used to produce *M. mycoides* JCVI-syn1.0. (b) Assembly of JCVI-syn1.0 genome occurred in three steps: (1) 1080-bp segments (cassettes; orange arrows), produced from overlapping synthetic oligonucleotides, were recombined in sets of 10 to produce 109 ~10-kb assemblies (blue arrows). (2) The 109 were then combined in sets of 10 to produce 11 ~100-kb assemblies (green arrows). (3) In the final step, the 11 segments were recombined to create the entire synthetic genome (red circle; the locations of *Asc I* and *BssH II* restriction sites are shown). All recombination steps were carried out in yeast. (c) Colonies of *M. mycoides* JCVI-syn1.0 (top) and wild-type *M. mycoides* (bottom). Cells were cultured on agar containing the compound X-gal. Colonies with the synthetic genome are blue because their cells contain the *lacZ* gene and express β-galactosidase, which metabolized X-gal to a blue compound (see Chapter 20). Wild-type cells do not express the *lacZ* gene and therefore remained white in color.

11 separate 100-kb assemblies that were eventually combined to completely span the entire 1.08-Mb *M. mycoides* genome [Figure 22.14(b)].

The entire assembled genome, called JCVI-syn1.0, was then transplanted into a close relative, *Mycoplasma capricolum*, as recipient cells. **Genome transplantation** is effectively the true test of the functionality of a synthetic genome. Transplanting an entire synthetic genome with the expectation that genes in the synthetic genome would completely transform the phenotype of the recipient cells is an essential outcome. Transplantation resulted in cells with the JCVI-syn1.0 genotype and the phenotype of a new strain of *M. mycoides* [**Figure 22.14(c)**]. Transformation of *M. capricolum* into JCVI-syn1.0 *M. mycoides* was verified, in part, because these cells were shown to express the *lacZ* gene, which was only present in the synthetic genome. Selection for tetracycline resistance and a determination that recipient cells also made proteins characteristic of *M. mycoides*, and not *M. capricolum*, were also used to verify strain conversion.

One particularly impressive accomplishment of these experiments was that the synthetic DNA was "naked" DNA, meaning it did not contain any proteins from *M. mycoides*. From this synthetic genome the cell was capable of transcribing all the appropriate genes and translating all the protein products necessary for life as *M. mycoides*. This is not a trivial accomplishment. The synthetic genome effectively rebooted the *M. capricolum* recipient cells to change them from one form to another. When this work was announced, J. Craig Venter claimed: "This is equivalent to changing a Macintosh computer into a PC by inserting a new piece of software."

This was tedious work, spanning over 15 years. Ninety-nine percent of the experiments involved failed! Keep in mind also, that these experiments did not create life from an inanimate object since they were based on converting one living strain into another. But clearly these studies provided key "proof of concept" that synthetic genomes could be produced, assembled, and successfully transplanted to create a microbial strain encoded by a synthetic genome. Thus this research brought scientists closer to producing novel synthetic genomes incorporating genes for specific traits of interest. Yet this work still did not define the minimal genome and answer the question of how simple the genome can be.

In 2016, JCVI announced that 473 genes is the minimal bacterial genome. To make this determination, JCVI created a synthetic version of the *M. mycoides* genome that was about half the size of JCVI-syn1.0 discussed earlier. This synthetic 531-kb genome, called JCVI-syn3.0, contained 473 genes, encoding 438 proteins and 35 RNAs. Using JCVI-syn3.0 and applying genome transplantation experiments resulted in colonies that are similar to those produced by JCVI-syn1.0 transplantation. About one-third (149 genes) of the genes in JCVI-syn3.0 that are essential for life have no known function. And several of these 149 genes are present in other organisms including humans. Investigating their roles is of significant interest to the JCVI team.

In the future, gene editing approaches such as CRISPR-Cas are expected to make it easier to alter genomes to address the minimal genome question. Several teams have already applied CRISPR-based functional analysis to bacteria such as *Bacillus subtilis*. Using CRISPR to knock out bacterial genes, geneticists can then screen for phenotypic changes as a way to identify essential genes in bacteria.

The Essential Genes of Human Cells and the Quest to Create a Synthetic Human Genome

Addressing questions about the minimal genome and the identification of essential genes in bacteria led scientists to explore answers to the same questions in more advanced eukaryotes, including yeasts and humans. The JCVI synthetic genome projects inspired researchers to create synthetic versions of six chromosomes, a little more than one-third, of the genome for *Saccharomyces cerevisiae*, a eukaryotic single-celled yeast.

The yeast work demonstrated that creating eukaryotic synthetic genomes is possible, but geneticists interested in the genes essential for a functional self-replicating human cell did not have the necessary tools to make significant progress experimentally. Thus, when the Human Genome Project was completed, creating a complete synthetic human genome was considered technically impossible. But once again, development of the CRISPR-Cas gene editing system has opened the door for scientists. Independently, three different groups recently obtained very similar results involving the core set of essential genes required for human cell division.

One approach involved using CRISPR-Cas to edit human genes from haploid cell lines derived from chronic myelogenous leukemia. Cell viability and RNA expression in edited cells were determined, among other parameters. Another group used a CRISPR gene editing strategy across many different cell types including cancer cells. These studies concluded that approximately 2000 genes, or about 10 percent of the 20,000 genes in the human genome, are the core set of essential genes required for human cell survival and replication. Human essential genes are also highly conserved across these cell lines, and many encode proteins that are critical for cellular protein—protein interactions.

The systematic assessment of all genes within a genome through gene editing by CRISPR-Cas provides yet another example of the value of this technique. Scientists are optimistic that studies such as these may also reveal novel insight about essential genes in cancer cells that can be used to develop novel strategies for disease treatment. There is much more work to be done on this topic, and it will be a rapidly developing area of human genome research to follow in the future.

In 2016 a group of synthetic biologists and genome scientists organized to propose an initiative, called the *Human Genome Project-write (HGP-write)*, to synthesize an entire human genome. Many have questioned the objectives for synthesizing a human genome and the ethical implications of doing so. With an estimated 10-year timeline and a cost of approximately $100 million, organizers of HGP-write hope this project will lead to advances in DNA synthesis technology at lower costs, in the same way that the Human Genome Project resulted in advances in DNA sequencing and dramatic reductions in the costs of sequencing an entire genome. Scientists associated with HGP-write say that another intention of this project is to eventually engineer genomes for model organisms and for human cells to help study disease—but not to engineer genomes for people.

Synthetic Biology for Bioengineering Applications

Venter's work with *M. mycoides* JCVI-syn1.0, a decade-long project that cost about $40 million, was hailed as a defining moment in the emerging field of **synthetic biology,** a field that applies engineering design principles to biological systems.

What are other potential applications of synthetic genomes and synthetic biology? One of JCVI's goals is to create microorganisms that can be used to synthesize biofuels. Other possibilities include creating synthetic microbes with genomes engineered to (1) express gene products to degrade pollutants (bioremediation); (2) synthesize new biopharmaceutical products; (3) synthesize chemicals and fuels from sunlight and carbon dioxide; and (4) produce "semisynthetic" crops that contain synthetic chromosomes encoding genes for beneficial traits such as drought resistance or improved photosynthetic efficiency.

Because of the complexity of creating an entire genome, synthetic biologists are using *bioengineering* approaches to put together parts of genes in different ways. Consider the components of a gene such as protein-coding sequences, start and stop triplets, promoters, and other regulatory sequences. Can these gene parts be interchanged and combined in unique ways, essentially reconstituting biological functions, for novel and valuable bioengineering applications?

In many ways this approach is borrowed from computer science. The parts of a circuit board and board design transformed computing. In synthetic biology, cells are DNA-driven machines where DNA sequences provide the software code that directs proteins to do work for cells. The International Genetically Engineered Machine (iGEM) Foundation maintains a Registry of Standard Biological Parts (see http://parts.igem.org/Main_Page) with thousands of working parts or interchangeable "components" for synthetic biology. Some believe that synthetic biology will transform applications of biology in the future. Even if novel applications develop slowly, synthetic biology will help us develop a deeper understanding of how the parts of a biological system work, such as component segments of a gene (regulatory sequences including promoters, enhancers, coding and noncoding sequences, and termination sequences).

George Church of Harvard University is one of the leaders in this field. His research team has created a synthetic gene circuit in *E. coli* by linking a series of three genes in a row. This system uses recombinase enzymes to cut stretches of DNA that flip the pieces and insert them back into the genome. Each flip of a DNA sequence is essentially like flipping an on or off switch, like a string of 0s and 1s in the binary code in a computer program. If this system is connected to a promoter sequence for a gene, each time the gene is turned on, the recombinases flip a section of DNA. Because the DNA sequence of bacterial cells is altered each time recombinase flips a piece of DNA, these sequence changes create a memory of the number of times a gene was turned on that can be determined by sequence analysis of flipped regions. Thus a DNA circuit has a memory for counting events. Such gene circuits have also been coupled to genes producing bioluminescent proteins as a way to report gene circuit activity (see **Figure 22.15**).

Church's team has connected their recombinase system to target sequences around the promoter for a gene controlling the production of green fluorescent protein

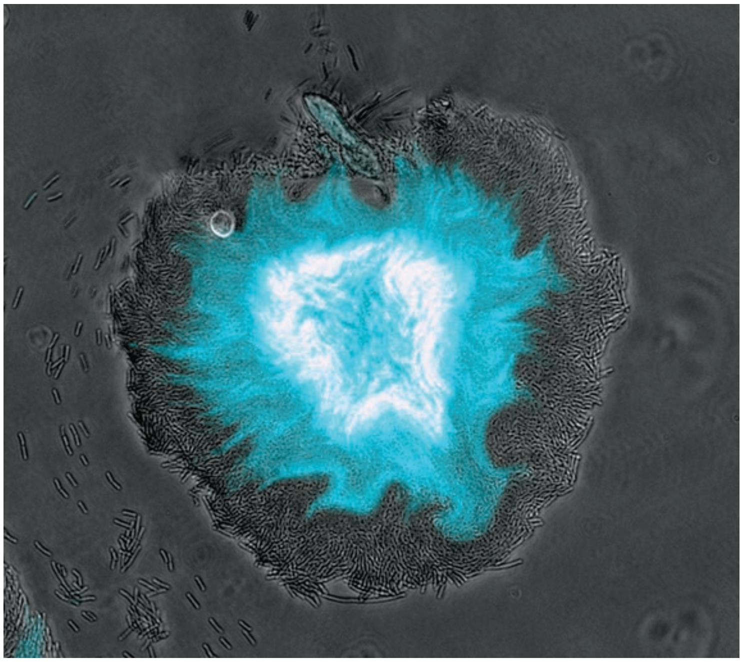


FIGURE 22.15 Gene circuits. In response to stimuli, colonies of bacteria engineered by synthetic biology produce synchronized flashes of bioluminescence under the control of a three-gene circuit.

(GFP). Each time this system is turned on, recombinases flip DNA sequences that turn on the promoter for GFP, cells express GFP, and they glow. The team can measure the number of inputs (turning on gene expression) by either sequencing the DNA or by monitoring GFP fluorescence from the cells. Computer scientists refer to these properties as logic functions and memory, in which specific inputs cause a program to respond (logic) and the program retains a memory.

As another interesting example of synthetic biology, Church and other geneticists are reengineering the genetic code by creating new codon combinations and thus *genetically recoded organisms* (GROs). Recoding, or repurposing codons for enhancing genomes with functions not normally present in nature is an approach that may be used to incorporate novel amino acids into proteins to change the chemical properties of particular proteins. For example, one application of this technology could include making proteins resistant to degradation or providing them with stability under harsh conditions of temperature or pH, which can be valuable for novel commercial applications of recombinant proteins.

Recently Church's group designed a synthetic genome in *E. coli* in which all 62,214 instances (~5 percent of all *E. coli* codons) of seven different codons were replaced with a synonymous codon resulting in a synthetic genome incorporating only 57 of the 64 possible codons. For example, in the wild-type *E. coli* genome, the codon sequence AGU is the most frequently prevalent codon encoding the amino acid serine. The triplet (TCA) encoding the codon AGU was computationally identified throughout the *E. coli* genome and replaced in the synthetic genome to produce the codon UCA—a synonymous codon which also codes for serine. In this experiment approximately 91 percent of tested essential genes retained their functionality, thus demonstrating that proof of concept that codon replacement may be possible in synthetic genomes.

Another group of researchers have demonstrated computationally that adding a newly designed artificial base pair to a genome with the traditional four nucleotides, thus expanding RNA to six nucleotides, could in turn create proteins built from as many as 172 amino acids instead of the existing 20 amino acids. This approach may have potential applications for incorporating additional artificial amino acids into proteins and could provide a powerful breakthrough for drug developers.

Investments in new biotechnology companies focused on synthetic biology have increased dramatically in the past three years. It will be very interesting to watch the development of this field in the near future to see if synthetic biology approaches can generate novel, successful, and profitable applications.

22.9 Genetic Engineering, Genomics, and Biotechnology Raise Ethical, Social, and Legal Questions

Geneticists use recombinant DNA and genomic technologies to identify genes, diagnose and treat genetic disorders, produce commercial and pharmaceutical products, and solve crimes. However, the applications that arise from these technologies raise important ethical, social, and legal issues that must be identified, debated, and resolved. Here we present a brief overview of some current ethical debates concerning the uses of genetic technologies. (Please note that ethical issues associated with the CRISPR-Cas gene editing system and with gene therapy are discussed in Special Topic Chapter 1—CRISPR-Cas and Genome Editing, and Special Topic Chapter 5—Gene Therapy, respectively.)

Genetic Testing and Ethical Dilemmas

When the Human Genome Project was first discussed, scientists and the general public voiced concerns about how genome information would be used and how the interests of both individuals and society can be protected. To address these concerns, the **Ethical, Legal, and Social Implications (ELSI) Program** was established by the National Human Genome Research Institute [a division of the National Institutes of Health (NIH)]. The ELSI Program focuses on four areas: (1) privacy and fairness in the use and interpretation of genetic information, (2) the transfer of genetic knowledge from the research laboratory to clinical practice, (3) ways to ensure that participants in genetic research know and understand the potential risks and benefits of their participation and give informed consent, and (4) enhancement of public and professional education.

The majority of the most widely applied genetic tests that have been used to date have provided patients and physicians with information that improve quality of life. One example involves prenatal testing for phenylketonuria (PKU) and implementing dietary restrictions to diminish the effects of the disease. But many of the potential benefits and consequences of genetic testing are not always clear. For example,

- We have the technologies to test for genetic diseases for which there are no effective treatments. *Should we test people for these disorders?*
- With current genetic tests, a negative result does not necessarily rule out future development of a disease, nor does a positive result always mean that an individual will get the disease. *How can we effectively communicate the results of testing and the actual risks to those being tested?*

- *What information should people have before deciding to have a genome scan or a genetic test for a single disorder or to have their whole genome sequenced?*
- Sequencing fetal genomes from the maternal bloodstream has revealed examples of mutations in the fetal genome (for example, a gene involved in Parkinson disease). *How might parents and physicians use this information?*
- Because sharing patient data through electronic medical records is a significant concern, *what issues of consent need to be considered?*
- *How can we protect the information revealed by genetic tests?*
- *How can we define and prevent genetic discrimination?*

Let's consider a specific case. In 2011, a case in Boston revealed the dangers of misleading results based on genetic testing. A prenatal ultrasound of a pregnant woman revealed a potentially debilitating problem (Noonan syndrome) involving the spinal cord of the woman's developing fetus. Physicians ordered a DNA test, which came back positive for a gene variant in a database that listed the gene as implicated in Noonan syndrome. The parents chose to terminate the pregnancy. Months later it was learned that the locus linked to Noonan was not involved in the disease, yet there was no effective way to inform the research and commercial genetic testing community.

To minimize these kinds of problems in the future, the NIH National Center for Biotechnology Information (NCBI) has developed a database called ClinVar (see www.ncbi.nlm.nih.gov/clinvar/) which integrates data from clinical genetic testing labs and research literature to provide an updated resource for researchers and physicians.

Disclosure of *incidental results* is another ethically challenging issue. When someone has his or her genome sequenced or has a test done involving a particular locus thought to be involved in a disease condition, the analysis sometimes reveals other mutations that could be of significance to the patient. Researchers and clinicians are divided on whether such information should be disclosed to the patient or whether patients should be asked for consent to receive all results from such tests. For example, a recent study considered 26 pregnant women who underwent prenatal genetic testing and learned they had genes associated with certain cancers and cognitive disorders as well as sex-chromosome abnormalities associated with reduced fertility. Again, this raises ethical issues about what type of consent women should consider when having these tests. Should these results be disclosed to these women? What do you think?

Earlier in this chapter we discussed preimplantation genetic diagnosis (PGD), which provides couples with the ability to screen embryos created by *in vitro* fertilization for genetic diseases. As we learn more about genes involved in human traits, will other, non-disease-related genes be screened for by PGD? Will couples be able to select embryos with certain genes encoding desirable traits for height, weight, intellect, or other physical or mental characteristics? What do you think of using genetic testing to purposely select for an embryo with a genetic disorder? There have been several well-publicized cases of couples seeking to use prenatal diagnosis or PGD to select for embryos with dwarfism and deafness.

As identification of genetic traits becomes more routine in clinical settings, physicians will need to ensure genetic privacy for their patients. There are significant concerns about how genetic information could be used in negative ways by employers, insurance companies, governmental agencies, or the general public. Genetic privacy and prevention of genetic discrimination will be increasingly important in the coming years. In 2008, the **Genetic Information Nondiscrimination Act (GINA)** was signed into law in the United States. This legislation is designed to prohibit the improper use of genetic information in health insurance and employment, but not life insurance.

Direct-to-Consumer Genetic Testing and Regulating the Genetic Test Providers

The past decade has seen dramatic developments in **direct-to-consumer (DTC) genetic tests.** A simple Web search will reveal many companies offering DTC genetic tests. As of 2017, there were over 2000 diseases for which such tests are now available (in 1993 there were about 100 such tests). Most DTC tests require that a person mail a saliva sample, hair sample, or cheek cell swab to the company. For a range of pricing options, DTC testing companies largely use SNP-based tests such as ASO tests to screen for different mutations. For example, in 2007, Myriad Genetics, Inc., began a major DTC marketing campaign of its tests for *BRCA1* and *BRCA2*. Mutations in these genes increase the risk of developing breast and ovarian cancer. DTC testing companies report absolute risk, the probability that an individual will develop a disease, but how such risks are calculated is highly variable and subject to certain assumptions.

Such tests are controversial for many reasons. For example, the test is purchased online by individual consumers and requires no involvement of a physician or other health-care professionals such as a nurse to administer the test or a genetic counselor to interpret the results. There are significant questions about the quality, effectiveness, and accuracy of such products because currently the DTC testing industry is largely self-regulated. The FDA does not regulate DTC genetic tests. There is at present no comprehensive way for patients to make comparisons and evaluations about the range of tests available and their relative quality.

Most companies make it clear that they are not trying to diagnose or prevent disease, nor are they offering health advice, so what is the purpose of the information that test results provide to the consumer? Web sites and online programs from DTC testing companies provide information on what advice a person should pursue if positive results are obtained. But is this enough? If results are not understood, might negative tests not provide a false sense of security? Just because a woman is negative for *BRCA1* and *BRCA2* mutations *does not* mean that one cannot develop breast or ovarian cancer. Refer to end-of-chapter Problem 17 for an example of a personal decision that actress Angelina Jolie made based on the results of a genetic test.

Whether the FDA will oversee DTC genetic tests in the future is unclear. However, at the time of publication of this edition, the FDA has not revealed any definitive plans to regulate or oversee DTC genetic tests. But because some DTC genetic testing companies, such as 23andMe, offer health-related analyses or health reports, they do fall under FDA regulations. The FDA continues to issue warnings to DTC testing companies to provide what the FDA considers to be appropriate health-related interpretations of genetic tests. For example, in 2017 for the first time the FDA approved a DTC saliva-test from 23andMe that can test for genetic mutations associated with 10 conditions including Parkinson disease and Alzheimer disease. There are varying opinions on the regulatory issue. Some believe that the FDA has no business regulating DTC tests and that consumers should be free to purchase products based on their personal needs or interests. Others insist that the FDA must regulate DTCs in the interest of protecting consumers.

DNA and Gene Patents

Intellectual property (IP) rights are being debated as an aspect of the ethical implications of genetic engineering, genomics, and biotechnology. Patents on intellectual property (isolated genes, new gene constructs, recombinant cell types, GMOs) can be potentially lucrative for the patent-holders but may also pose ethical and scientific problems.

Why is protecting IP important for companies? Consider this issue. If a company is willing to spend millions or billions of dollars and several years doing research and development (R&D) to produce a valuable product, then shouldn't it be afforded a period of time to protect its discovery so that it can recover R&D costs and made a profit on its product?

Genes in their natural state as products of nature cannot be patented. Consider the possibilities for a human gene that has been cloned and then patented by the scientists who did the cloning. The person or company holding the patent could require that anyone attempting to do research with the patented gene pay a licensing fee for its use. Should a diagnostic test or therapy result from the research, more fees and royalties may be demanded, and as a result the costs of a genetic test may be too high for many patients to afford. But limiting or preventing the holding of patents for genes or genetic tools could reduce the incentive for pursuing the research that produces such genes and tools, especially for companies that need to profit from their research.

Should scientists and companies be allowed to patent DNA sequences from naturally living organisms? And should there be a lower or an upper limit to the size of those sequences? For example, should patents be awarded for small pieces of genes, just because some individual or company wants to claim a stake in having cloned that specific sequence of DNA first, even if no one knows whether the DNA sequence has a use? Can or should investigators be allowed to patent the entire genome of any organism they have sequenced?

It is estimated that the U.S. Patent and Trademark Office has granted patents for approximately 20 percent of the genes in the human genome. Incidentally the patenting of human genes has led some to use the term *patentome*! Some scientists are concerned that to award a patent for simply cloning a piece of DNA is awarding a patent for too little work. Given that computers do most of the routine work of genome sequencing, who should get the patent? What about individuals who figure out *what* to do with the gene? What if a gene sequence has a role in a disease for which a genetic therapy may be developed? Many scientists believe that it is more appropriate to patent novel technology and applications that make use of gene sequences than to patent the gene sequences themselves.

In recent years, the Supreme Court has ruled on cases related to patenting of human genes and any sequences, functions, or correlations to naturally occurring products from a gene. Patenting of genetic tests is also under increased scrutiny in part because of concerns that a patented test can create monopolies in which patients cannot get a second opinion if only one company holds the rights to conduct a particular genetic test. Recent analysis has estimated that as many as 64 percent of patented tests for disease genes make it very difficult or impossible for other groups to propose a different way to test for the same disease.

In 2010 a landmark case brought by the American Civil Liberties Union against Myriad Genetics contended that Myriad could not patent the *BRCA1* and *BRCA2* gene sequences used to diagnose breast cancer. Myriad's BRACAnalysis® product has been used to screen over a million women for *BRCA1* and *BRCA2* during its period of patent exclusivity. A U.S. District Court judge ruled Myriad's patents invalid on the basis that DNA in an isolated form is not fundamentally different from how it exists in the body. Myriad was essentially accused of having a monopoly on its tests, which have existed for a little over a decade based on its exclusive licenses in the United States.

This case went to the Supreme Court in 2013, which rendered a 9—0 ruling against Myriad, stripping it of five of its patent claims for the *BRCA1* and *BRCA2* genes, largely based on the view that natural genes are a product of nature and just because they are isolated does not mean they can be patented. The Court ruled that cDNA sequences produced in a lab can continue to be patentable. Myriad still holds about 500 valid claims related to *BRCA* gene testing.

Whole-Genome Sequence Analysis Presents Many Questions of Ethics

In the next decade and beyond, it is expected that WGS analysis of adults and babies will become increasingly common in clinical settings. A Genomic Sequencing and Newborn Screening Disorders (GSNSD) research program initiated by the NIH is under way to sequence the exomes and genomes of more than 1500 babies. Both infants with illnesses and babies who are healthy will be part of this study. This NIH-funded initiative will allow scientists to carry out comparative genomic analyses of specific sequences to help identify genes involved in disease conditions.

Screening of newborns is important to help prevent or minimize the impacts of certain disorders. Each year routine blood tests from a heel prick of newborn babies reveal rare genetic conditions in several thousand infants in the United States alone. A small number of states allow parents to opt out of newborn testing. In the future, should DNA sequencing at the time of birth be universally required? Do we really know enough about which human genes are involved in disease to help prevent disease in children? Estimates suggest that sequencing can identify approximately 15 to 50 percent of children with diseases that currently cannot be diagnosed by other methods. What is the value of having sequencing data for healthy children?

Personal genomics adds another layer of complexity to this discussion. When people donate their DNA for WGS projects, should they have access to the raw data from their sequence analysis? Currently, such volunteers are refused access to their genetic data.

As exciting as this period of human genetics and medicine is becoming, most of the WGS studies of individuals are happening in a largely unregulated environment. This raises significant ethical concerns especially with respect to DNA collection, the variability and quality control of DNA handling protocols, sequence analysis, storage, and confidentiality of genetic information (see the Genetics, Ethics, and Society box).

GENETICS, ETHICS, AND SOCIETY

Privacy and Anonymity in the Era of Genomic Big Data

Our lives are surrounded by Big Data. Enormous quantities of personal information are stored on private and public databases, revealing our purchasing preferences, search engine histories, social contacts, and even GPS locations. But often we do not know how this information may be used in the future, or how its distribution might affect us, our families, and our relationships.

Perhaps the most personal of all Big Data entries are those obtained from personal genome sequences and genomic analyses. Tens of thousands of individuals are now donating DNA for whole-genome sequencing—by both private gene-sequencing companies and public research projects. Most people who donate their DNA for sequence analysis do so with little concern. After all, what consequences could possibly come from access to gigabytes of A's, C's, T's, and G's? Surprisingly, the answer is—quite a lot.

One of the first inklings of genetic privacy problems arose in 2005, when a 15-year-old boy named Ryan Kramer tracked down his anonymous sperm-donor father using his own Y chromosome sequence data and the Internet [Motluk, A. (2005) *New Scientist* 2524: November 3]. Ryan submitted a DNA sample to a genealogy company that generates Y chromosome profiles, matches them against entries in their database, and puts people into contact with others who share similar genetic profiles, indicating relatedness. Two men contacted Ryan, and both had the same last name, with different spellings. Ryan combined the information about his potential relatives' last names with the only other information that he had about his sperm-donor father—date of birth, birth place, and college degree. Using an Internet search, he obtained the names of everyone born on that date in that place. On the list, there was one man with the same last name as his two Y chromosome relatives. He confirmed that the man also had the appropriate college

(continued)

Genetics, Ethics, and Society, continued

degree and then contacted his sperm-donor father. Since this report, other children of sperm donors have used DNA genealogy searching to find their paternal parent. The implications for sperm donors have been unsettling, as most are promised anonymity. In some cases, donors are troubled to learn that they have fathered dozens of offspring.

More recently, several published reports reveal the ease with which anyone's identity can be traced using DNA-sequence profiles and Internet searches. These searches can reveal people's identities and other personal information such as age, sex, body mass index, glucose, insulin, lipid levels, and disease susceptibilities.

To many people, the implications of "genomic re-identification" are disturbing. Genomic information leaks could reveal personal medical information, physical appearance, and racial origins. They could also be used to synthesize DNA to plant at a crime scene or could be used in unforeseen ways in the future as we gain more information about what resides in our genome. The consequences of genomic information leaks also encompass family members from many generations, who share the person's genetic heritage.

Your Turn

Take time, in pairs or in larger groups, to consider the following questions and references concerning the ethical and technical challenges of ensuring genetic privacy.

1. What are some of the ethical arguments for and against maintaining genetic privacy and anonymity?

A discussion of ways to balance the need for privacy with the need for research information is presented in Hansson, M. G., et al. (2016). The risk of re-identification versus the need to identify individuals in rare disease research. *Eur. J. Hum. Genet.* 24:1553–1558.

1. Would you be willing to send a DNA sample to a private company for whole-genome sequencing? If not, why not? If so, what privacy assurances would you need to make you comfortable about ordering your genome sequence?

For information on privacy policies of direct-to-consumer DNA-sequencing companies, see Niemiec, E., and Howard, H. C. (2016). Ethical issues in consumer genome sequencing: Use of consumers' samples and data. *Appl. Transl. Genom.* 8:23–30.

Preconception Testing, Destiny Predictions, and Baby-Predicting Patents

Companies are now promoting the ability to do *preconception* testing and thus make "destiny predictions" about the potential phenotypes of hypothetical offspring based on computational methods for analyzing sequence data of parental DNA samples. The company 23andMe has been awarded a U.S. patent for a computational method called the *Family Traits Inheritance Calculator* to use parental DNA samples to predict a baby's traits, including eye color and the risk of certain diseases. This patent includes applications of technologies to screen sperm and ova for IVF.

Currently, gender selection of embryos generated by IVF is very common. But could preconception testing lead to the selection of "designer babies"? Fear of *eugenics* surrounds these conversations, particularly as genetic analysis starts moving away from disease conditions to nonmedical traits such as hair color, eye color, other physical traits, and potentially behavioral traits. The patent has been awarded for a process that will compare the genotypic data of an egg provider and a sperm provider to suggest gamete donors that might result in a baby or hypothetical offspring with particular phenotypes of interest to a prospective parent. What do you think about this?

A company called GenePeeks claims to have a patent-pending technology for reducing the risk of inherited disorders by "digitally weaving" together the DNA of prospective parents. GenePeeks plans to sequence the DNA of sperm donors and women who want to get pregnant to inform women about donors who are most genetically compatible for the traits they seek in offspring. Their proprietary computing technology is intended to use sequence data to examine "virtual" eggs and sperm from donor—client pairings to estimate the likelihood of about 10,000 specific diseases in hypothetical offspring from prospective parents. Will technologies such as this become widespread and attract consumer demand in the future? What do you think? Would you want this analysis done before deciding whether to have a child with a particular person?

Patents and Synthetic Biology

The J. Craig Venter Institute (JCVI) has filed patent applications for what is being called "the world's first-ever human-made life form." The patents are intended to cover the minimal genome of *M. genitalium* (discussed in Section 22.8), which JCVI believes are the genes essential for self-replication. One of these patent applications is designed to claim the rights to synthetically constructed organisms. Another U.S. patent issued to a different group of researchers covers application of a minimal genome for *E. coli*, which has generated even more concern given its relative importance in research and commercial applications compared with *M. genitalium*. What do you think? Should it be possible to patent a minimal genome or a synthetic organism?

There are still other ethical issues about synthetic biology that merit consideration. Synthetic biology has the potential to be used for harmful purposes (such as bioterrorism). What regulatory policies and restrictions should be placed on applications of synthetic biology and on patents of these applications? The ability to modify life forms is offensive to many peoples' beliefs. How will this issue be addressed by the synthetic biology research community?

CASE STUDY “Driving” to Extinction

A new generation of methods for genome modification is emerging from research laboratories, extending the reach of biotechnology to many new fields. Although applications of these methods are still under development, their potential uses have already raised serious bioethical questions. One of these new technologies is a method for reducing or eliminating mosquito species that spread disease. For example, one species, *Aedes aegypti*, is involved in the death of more humans than any other animal. This insect carries and spreads malaria, yellow fever, dengue fever, chikungunya, and Zika fever. Malaria alone kills over 400,000 people each year. A technology called *gene drive* allows a desired allele to enter and spread throughout a population, until it has completely replaced an existing allele. In the case of *A. aegypti*, the allele of choice would cause female sterility and larval lethality. Gene drive could thus be used to drive this mosquito species to extinction in many areas, thereby reducing or eliminating all the diseases it carries.

1. Aside from the benefits of controlling the spread of serious diseases, there are ethical issues associated with the use of gene drive. What ethical issues should be considered before releasing a gene drive system into ecosystems?
2. What genetic controls or limitations should be put in place before releasing gene drive systems into wild populations?
3. Should gene drive systems be considered as tools to eliminate invasive species that cause widespread damage to ecosystems, such as the cane toad in Australia or the brown snakes on Guam?

For related reading, see Oye, K., et al. (2015). Regulating gene drives. *Science* 345:626–628.

Summary Points

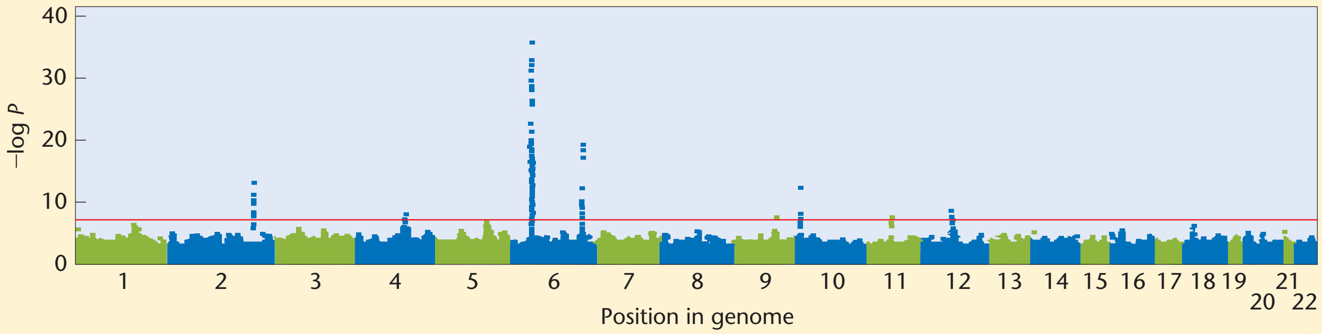
Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Recombinant DNA technology can be used to produce valuable biopharmaceutical protein products such as therapeutic proteins for treating disease.
2. Transgenic animals with improved growth characteristics or desirable phenotypes are being genetically engineered for a number of different applications.
3. A variety of different molecular techniques and genomic analyses, including allele-specific oligonucleotides tests, DNA microarrays, whole-genome sequencing, and RNA sequencing can be used to identify genotypes associated with both normal and disease phenotypes.
4. Preimplantation genetic diagnosis and noninvasive methods for deducing a fetal genome from maternal blood allow for genetic analysis of a developing fetus.
5. Whole-genome sequencing of individual genomes is entering medical clinics and becoming valuable for diagnosis and treatment of genetic conditions.
6. Genome-wide association studies can reveal genetic variations linked with disease conditions within populations.
7. Interest in synthetic biology and its potential applications has been spurred on by the creation of a synthetic genome successfully used in a genome transplantation experiment in bacteria.
8. Applications of genetic engineering and biotechnology involve a range of ethical, social, and legal dilemmas with important scientific and societal implications.

INSIGHTS AND SOLUTIONS

1. Research by Petukhova et al. [(2010) *Nature* 466:113–117] involved a genome-wide association study to analyze 1054 cases of patients with alopecia areata (AA) and 3278 controls. Alopecia areata is a condition that leads to major hair loss and affects approximately 5.3 million people in the United States alone. The study identified 139 single nucleotide polymorphisms significantly associated with AA.

   (a) A Manhattan plot from this work is shown below.



(continued)

Insights and Solutions—continued

Based on your interpretation of this plot, which chromosomes were associated with loci that may contribute to AA?

(b) Of the 139 SNPs significantly associated with AA, several are located within genes involved in controlling the activation and proliferation of regulatory T lymphocytes (T_{reg} cells) and cytotoxic T lymphocytes, genes involved in antigen presentation to immune cells, genes for immune regulatory molecules such as the interleukins, and genes expressed in the hair follicle itself. Speculate how these candidate genes may help scientists understand how AA progresses as a disease.

Solution:
(a) Investigators identified eight genomic regions with SNPs that exceed the genome-wide significance value of 5×10^{-7} (red line). These regions were clustered on chromosomes 2, 4, 6, 9, 10, 11, and 12.

(b) AA is an autoimmune disease in which the immune system attacks hair follicles, resulting in hair loss that can permeate across the entire scalp and even the whole body. AA hair follicles are attacked by T cells. The identification of candidate genes involved in T-cell proliferation, immune system regulation, and follicular development may potentially help investigators develop cures for AA.

2. Infection by HIV-1 (human immunodeficiency virus) weakens the immune system and results in the symptoms of AIDS (acquired immunodeficiency syndrome). Specifically, HIV infects and kills cells of the immune system that carry a cell-surface receptor known as CD4. An HIV surface protein known as gp120 binds to the CD4 receptor and allows the virus to enter the cell. The gene encoding the CD4 protein has been cloned. How might this clone be used along with recombinant DNA techniques to combat HIV infection?

Solution:
Researchers hope that clones of the *CD4* gene can be used in the design of systems for the targeted delivery of drugs and toxins to combat the infection. For example, because infection depends on an interaction between the viral gp120 protein and the CD4 protein, the cloned *CD4* gene has been modified to produce a soluble form of the protein (sCD4) that, because of its solubility, would circulate freely in the body. The idea is that HIV might be prevented from infecting cells if the gp120 protein of the virus first encounters and binds to extra molecules of the soluble form of the CD4 protein. Once bound to the extra molecules, the virus would be unable to bind to CD4 proteins on the surface of immune system cells. Studies in cell-culture systems indicate that the presence of sCD4 effectively prevents HIV infection of tissue culture cells. However, studies in HIV-positive humans have been somewhat disappointing, mainly because the strains of HIV used in the laboratory are different from those found in infected individuals.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on a number of interesting applications of genetic engineering, genomics, and biotechnology. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
(a) What experimental evidence confirms that we have introduced a useful gene into a transgenic organism and that it performs as we anticipate?
(b) How does a positive ASO test for sickle-cell anemia determine that an individual is homozygous recessive for the mutation that causes sickle-cell anemia?
(c) From microarray analysis how do we know what genes are being expressed in a specific tissue?
(d) How can we correlate the genome with RNA expression data in a tissue or a single cell?
(e) From GWAS how do we know which genes are associated with a particular genetic disorder?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 521. Most of these center on applications of genetic technology that are becoming widespread. Write a short essay that summarizes the impacts that genomic applications are having on society and discuss which of the ethical issues presented by these applications is the most daunting to society.

3. Why are most recombinant human proteins produced in animal or plant hosts instead of bacterial host cells?

4. One of the major causes of sickness, death, and economic loss in the cattle industry is *Mannheimia haemolytica*, which causes bovine pasteurellosis, or shipping fever. Noninvasive delivery of a vaccine using transgenic plants expressing immunogens would reduce labor costs and trauma to livestock. An early step toward developing an edible vaccine is to determine whether an injected version of an antigen (usually a derivative of the pathogen) is capable of stimulating the development of antibodies in a test organism. The following table assesses the ability of a transgenic portion of a toxin (Lkt) of *M. haemolytica* to stimulate development of specific antibodies in rabbits.

Immunogen Injected	Antibody Production in Serum
Lkt50*—saline extract	+
Lkt50*—column extract	+
Mock injection	−

*Lkt50 is a smaller derivative of Lkt that lacks all hydrophobic regions. + indicates at least 50 percent neutralization of toxicity of Lkt; − indicates no neutralization activity.

Source: Modified from Lee et al. (2001). *Infect. and Immunity* 69:5786–5793.

(a) What general conclusion can you draw from the data?
(b) With regards to development of a usable edible vaccine, what work remains to be done?

5. Sequencing the human genome, the development of microarray technology, and personal genomics promise to improve our understanding of normal and abnormal cell behavior. How are

these approaches dramatically changing our understanding and treatment of complex diseases such as cancer?

6. A couple with European ancestry seeks genetic counseling before having children because of a history of cystic fibrosis (CF) in the husband's family. ASO testing for CF reveals that the husband is heterozygous for the *Δ508* mutation and that the wife is heterozygous for the *R117* mutation. You are the couple's genetic counselor. When consulting with you, they express their conviction that they are not at risk for having an affected child because they each carry different mutations and cannot have a child who is homozygous for either mutation. What would you say to them?

7. As genetic testing becomes widespread, medical records will contain the results of such testing. Who should have access to this information? Should employers, potential employers, or insurance companies be allowed to have this information? Would you favor or oppose having the government establish and maintain a central database containing the results of individuals' genome scans?

8. Might it make sense someday to sequence every newborn's genome at the time of birth? What are the potential advantages and concerns of this approach?

9. Which of the examples of genetic testing below are prognostic tests? Which are diagnostic?

(a) Individual sequencing (personal genomics) identifies a mutation associated with Alzheimer's disease.

(b) ASO testing determines that an individual is a carrier for the mutant β-globin allele (β^S) found in sickle-cell anemia.

(c) DNA sequencing of a breast tumor reveals mutations in the *BRCA1* gene.

(d) Genetic testing in a healthy teenager identifies an SNP correlated with autism.

(e) An adult diagnosed with Asperger syndrome (AS) has a genetic test that reveals a SNP in the *GABRB3* gene that is significantly more common in people with AS than the general population.

10. Does genetic analysis by ASO testing allow for detection of epigenetic changes that may contribute to a genetic disorder? Explain your answer.

11. Maternal blood tests for three pregnant women revealed they would be having boys, yet subsequent ultrasound images showed all three were pregnant with girls. In each case Y chromosome sequences in each mother's blood originated from transplanted organs they had received from men! This demonstrates one dramatic example of a limitation of genetic analysis of maternal blood samples. What kind of information could have been collected from each mother in advance of these tests to better inform physicians prior to performing each test?

12. What is the main purpose of genome-wide association studies (GWAS)? How can information from GWAS be used to inform scientists and physicians about genetic diseases?

13. Describe how the team from the J. Craig Venter Institute created a synthetic genome. How did the team demonstrate that the genome converted the recipient strain of bacteria into a different strain?

14. Consider ethical issues associated with creating a synthetic human genome. Are there specific applications for a synthetic human genome that you support? Is creating a synthetic genome enhanced with genes for certain kinds of traits one of those applications?

15. The family of a sixth-grade boy in Palo Alto, California, was informed by school administrators that he would have to transfer out of his middle school because they believed his mutation of the *CFTR* gene, which does not produce any symptoms associated with cystic fibrosis, posed a risk to other students at the school who have cystic fibrosis. After missing 11 days of school, a settlement was reached to have the boy return to school. What ethical problems might you associate with this example?

16. Dominant mutations can be categorized according to whether they increase or decrease the overall activity of a gene or gene product. Although a loss-of-function mutation (a mutation that inactivates the gene product) is usually recessive, for some genes, one dose of the normal gene product, encoded by the normal allele, is not sufficient to produce a normal phenotype. In this case, a loss-of-function mutation in the gene will be dominant, and the gene is said to be *haploinsufficient*. A second category of dominant mutation is the gain-of-function mutation, which results in a new activity or increased activity or expression of a gene or gene product. The gene therapy technique currently being used in clinical trials involves the "addition" to somatic cells of a normal copy of a gene. In other words, a normal copy of the gene is inserted into the genome of the mutant somatic cell, but the mutated copy of the gene is not removed or replaced. Will this strategy work for either of the two aforementioned types of dominant mutations?

17. In 2013 the actress Angelina Jolie elected to have prophylactic double-mastectomy surgery to prevent breast cancer based on a positive test for mutation of the *BRCA1* gene. What are some potential positive and negative consequences of this high-profile example of acting on the results of a genetic test?

18. The National Institutes of Health created the *Genetic Testing Registry (GTR)* to increase transparency by publicly sharing information about the utility of their tests, research for the general public, patients, health-care workers, genetic counselors, insurance companies, and others. The Registry is intended to provide better information to patients, but companies involved in genetic testing are not required to participate. Should company participation be mandatory? Why or why not? Explain your answers.

19. Should the FDA regulate direct-to-consumer genetic tests, or should these tests be available as a "buyer beware" product?

Extra-Spicy Problems

20. Would you have your genome sequenced, if the price was affordable? Why or why not? If you answered yes, would you make your genome sequence publicly available? How might such information be misused?

21. Following the tragic shooting of 20 children at a school in Newtown, Connecticut, in 2012, Connecticut's state medical examiner requested a full genetic analysis of the killer's genome. What do you think investigators might be looking for? What might they expect to find? Might this analysis lead to oversimplified analysis of the cause of the tragedy?

22. Private companies are offering personal DNA sequencing along with interpretation. What services do they offer? Do you think that these services should be regulated, and if so, in what way? Investigate one such company, 23andMe, at http://www.23andMe.com, before answering these questions.

23. Yeager, M., et al. [(2007) *Nature Genetics* 39:645–649] and Sladek, R., et al. [(2007) *Nature* 445:881–885] have used single-nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS) to identify novel risk loci for prostate cancer and Type 2 diabetes, respectively. Each study suggests that disease-risk genes can be identified that significantly contribute to the disease state. Given your understanding of such complex diseases, what would you determine as reasonable factors to consider when interpreting the results of GWAS?

24. In 2010, a U.S. District Judge ruled to invalidate Myriad Genetics' patents on the *BRCA1* and *BRCA2* genes. Judge Sweet noted that since the genes are part of the natural world, they are not patentable. Myriad Genetics also holds patents on the development of a direct-to-consumer test for the *BRCA1* and *BRCA2* genes.

(a) Would you agree with the ruling to invalidate the patenting of the *BRCA1* and *BRCA2* genes? If you were asked to judge the patenting of the direct-to-consumer test for the *BRCA1* and *BRCA2* genes, how would you rule?

(b) J. Craig Venter has filed a patent application for his "first-ever human-made life form." This patent is designed to cover the genome of *M. genitalium*. Would your ruling for Venter's "organism" be different from the judge's ruling on patenting of the *BRCA1* and *BRCA2* genes?

25. A number of mouse models for human cystic fibrosis (CF) exist. Each of these mouse strains is transgenic and bears a different specific *CFTR* gene mutation. The mutations are the same as those seen in several varieties of human CF. These transgenic CF mice are being used to study the range of different phenotypes that characterize CF in humans. They are also used as models to test potential CF drugs. Unfortunately, most transgenic mouse CF strains do not show one of the most characteristic symptoms of human CF, that of lung congestion. Can you think of a reason why mouse CF strains do not display this symptom of human CF?

26. Craig Venter and others have constructed synthetic copies of viral genomes. For example, the genome for poliovirus and the 1918 influenza strain responsible for the pandemic flu have been assembled this way. The United States currently has a moratorium on federal funding for "gain-of-function" experiments which increase the virulence or transmission potential of viruses. What concerns might ethicists have about synthetic biology studies involving potential pandemic pathogens?

23

Developmental Genetics

This unusual four-winged *Drosophila* has developed an extra set of wings as a result of a mutation in a homeotic selector gene.

CHAPTER CONCEPTS

- Development is a process by which cells undergo progressive stages of structural and functional specialization as a result of differential gene expression.
- Animals use a small number of shared signaling systems and regulatory networks to construct a wide range of adult body forms from the zygote. These shared properties make it possible to use animal models to study human development.
- Differentiation is controlled by cascades of gene expression that are a consequence of events that specify and determine the developmental fate of cells.
- Plants independently evolved developmental regulatory mechanisms that parallel those of animals.
- Cell–cell signaling programs the developmental fate of adjacent as well as distant cells.
- In many organisms, binary switch genes program the developmental fate of embryonic cells.

Over the last two decades, genetic analysis, molecular biology, and genomics have shown that, in spite of wide diversity in the size and shape of adult animals and plants, multicellular organisms share many genes, genetic pathways, and molecular signaling mechanisms that control developmental events leading from the zygote to the adult. At the cellular level, development is marked by three important events: **specification,** when the first cues confer spatially distinct identity; **determination,** the time when a specific developmental fate for a cell becomes fixed; and **differentiation,** the process by which a cell achieves its final adult form and function. Thanks to newly developed methods of analysis including microarrays, high-throughput sequencing, epigenetics, proteomics, and systems biology, we are beginning to understand how the expression and interaction of genes with environmental factors control developmental processes in eukaryotes.

In this chapter, the primary emphasis will be on how genetic analysis has been used to study development. This field, called developmental genetics, laid the foundation for our understanding of developmental events at the molecular and cellular levels, which contribute to the continually changing phenotype of the newly formed organism.

23.1 Differentiated States Develop from Coordinated Programs of Gene Expression

Animal genomes contain tens of thousands of genes, but only a small subset of these control the events that shape the adult body (Figure 23.1).

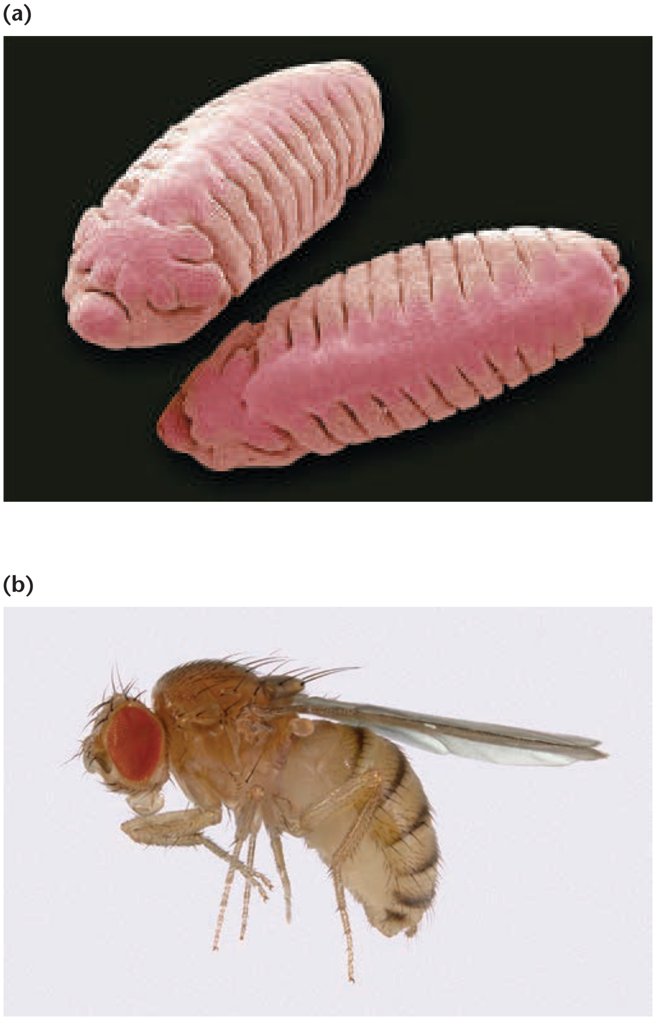


FIGURE 23.1 (a) A *Drosophila* embryo and (b) the adult fly that develops from it.

Developmental geneticists study mutant alleles of these genes to ask important questions about development:

- What genes are expressed?
- When are they expressed?
- In what parts of the developing embryo are they expressed?
- How is the expression of these genes regulated?
- What happens when these genes are defective?

These questions provide a foundation for exploring the molecular basis of developmental processes such as determination, induction, cell—cell communication, and cellular differentiation. Genetic analysis of mutant alleles is used to establish a causal relationship between the presence or absence of inducers, receptors, transcriptional events, cell and tissue interactions, and the observable structural changes that accompany development.

A useful way to define **development** is to say that it is the *attainment of a differentiated state* by all the cells of an organism (except for stem cells). For example, a cell in a blastula-stage embryo (when the embryo is just a ball of uniform-looking cells) is undifferentiated, while a pancreatic cell synthesizing insulin in the adult body is differentiated. How do cells get from the undifferentiated to the differentiated state? The process involves progressive activation of different groups of gene sets in different cells of the embryo. From a genetics perspective, one way of defining the different cell types that form during development in multicellular organisms is to identify and catalog the genes that are active in each cell type. In other words, development depends on patterns of differential gene expression.

The idea that differentiation is accomplished by activating and inactivating genes at different times and in different cell types is called the **variable gene activity hypothesis.** Its underlying assumptions are, first, that each cell contains an entire genome and, second, that differential transcription of selected genes controls the development and differentiation of each cell.

Genetic and Epigenetic Regulation of Development

In mammals, as in other organisms, each differentiated cell type in the adult has a distinct pattern of gene expression that sets it apart from all other cell types that compose the tissues and organ systems of the body. At fertilization, an egg and sperm fuse to form a single **totipotent** cell, the zygote. Totipotent cells have the capacity to differentiate into any of the specialized cells of the adult as well as any of the cells associated with the embryo, including the placenta. After several rounds of division, the embryo forms a blastula and the totipotent cells begin to specialize. A cluster of 30—40 **embryonic stem cells (ESCs)** form. These cells are **pluripotent** and can form any of the 200 or so different cell types found in adults (but not the placental cells or other embryo-specific cells). The developmental fate of stem cells is shaped by a series of steps, each of which progressively restricts their developmental potential. These stages of differentiation are controlled by a changing pattern of gene regulation and by epigenetic events that lock in these new transcription profiles to ensure they are stably transmitted during cell division and proliferation. Some of the factors that contribute to this process are summarized in Waddington's model of the epigenetic landscape (see Chapter 19).

Epigenetic regulation of the steps involved in specification, determination, and differentiation is accomplished through modification of chromatin structure via DNA methylation, histone tail modifications, and the actions of noncoding RNA. These modifications create new and heritable patterns of gene expression along with changes in chromosome topology and nuclear organization.

The epigenetic process begins early in embryogenesis with a global DNA demethylation that erases parental epigenetic marks (Figure 23.2) and contributes to the

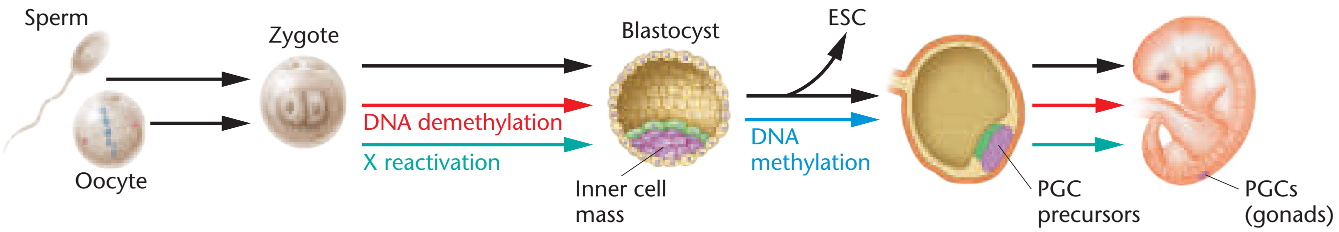


FIGURE 23.2 Epigenetic modifications of the genome help determine cell identity during development. After fertilization, global demethylation converts cells to a totipotent or pluripotent state. Two populations of pluripotent cells are created by a new cycle of DNA methylation in cells of the inner cell mass: embryonic stem cells (ESCs) and embryonic germ cells (EGCs). The ESCs will form the 200 different cell types in the body, and the EGCs will form primordial germ cells (PGCs), which migrate to the gonads and become precursors to sperm and eggs.

establishment of pluripotency. By the time the embryo reaches the blastula stage, methylation gradually resumes (Figure 23.2) and is associated with the initial steps in specification and determination of embryonic stem cells (ESCs) and **embryonic germ cells (EGCs).** The pluripotent ESCs will go on to form all the tissue types in the body, and the EGCs will form the primordial germ cells that will develop into gametes. As development proceeds, rounds of DNA methylation are accompanied by recruitment of chromatin-remodeling proteins that selectively activate and repress transcription. Regulatory genomic regions including enhancers and promoters are sites of DNA methylation and histone modification that activate or repress transcription. A study of histone modifications in fetal organs (heart, liver, and brain) shortly after their formation examined modifications at over 40,000 enhancers and identified both tissue-specific and developmental stage-specific alterations, some of which were in place to keep cells ready for subsequent developmental events.

In summary, epigenetic programming and reprogramming is important for establishing and maintaining cell identity during development. Erasure and reprogramming of epigenetic marks are normal parts of the mammalian life cycle, and knowledge of these epigenetic mechanisms will be important in developing methods for reprogramming somatic cells *in vitro* to create pluripotent cells that can be used in clinical cell replacement therapies.

23.2 Evolutionary Conservation of Developmental Mechanisms Can Be Studied Using Model Organisms

Genetic analysis of development across a wide range of organisms has shown that the size and shape of all animal bodies are controlled by a common set of genes and developmental mechanisms. For example, most of the differences in shape between zebras and zebrafish and thousands of other organisms are the result of different patterns of expression of highly conserved regulatory genes, such as the homeotic (abbreviated as *Hox*) genes, and not by species-specific genes. Genome-sequencing projects have confirmed that homeotic genes from a wide range of organisms have a common ancestry; this homology means that many aspects of normal human embryonic development and associated genetic disorders can be studied in model organisms such as *Drosophila melanogaster*, where genetic methods including mutagenesis, genetic crosses, and large-scale experiments involving hundreds of offspring can be conducted (see Chapter 1 for a discussion of model organisms in genetics).

Results from a field called evo-devo, which combines evolutionary and developmental biology, have revealed that although many developmental mechanisms are common to all animals, evolution has generated several new and unique ways of transforming a zygote into an adult. These evolutionary changes result from several genetic mechanisms including mutation, gene duplication and evolutionary divergence, the assignment of new functions to old genes, the recruitment of genes to new developmental pathways, and the modification of *cis*-regulatory sequences such as enhancers (see Chapter 17) that affect where and when during development regulatory genes are expressed. However, the emphasis in this chapter will be on the similarities in genes and developmental mechanisms among species.

Analysis of Developmental Mechanisms

In the space of this chapter, we cannot survey all aspects of development, nor can we explore in detail how developmental mechanisms triggered by the fusion of sperm and egg were identified. Instead, we will focus on a number of general processes in development:

- How the adult body plan of animals is laid down in the embryo
- The program of gene expression that turns undifferentiated cells into differentiated cells
- The role of cell–cell communication in development

We will use three model systems—the fruit fly *Drosophila melanogaster,* the flowering plant *Arabidopsis thaliana,* and the nematode *Caenorhabditis elegans*—to illustrate these developmental processes and related topics. We will examine how patterns of differential gene expression lead to the progressive restriction of developmental options resulting in the formation of the adult body plan in *Drosophila* and *Arabidopsis*. We will then expand the discussion to consider how our knowledge of events in these organisms has contributed to our understanding of developmental defects in humans. Finally, we will consider the role of cell–cell communication in the development of adult structures in *C. elegans*.

23.3 Genetic Analysis of Embryonic Development in *Drosophila* Reveals How the Body Axis of Animals Is Specified

How does a given cell at a precise position in the embryo switch on or switch off specific genes at timed stages of development? This is a central question in developmental biology. To answer this question, we will examine the sequence of gene expression in the embryo of the model organism *Drosophila*. Although development in a fruit fly appears to have little in common with humans, recall that the same genes regulate development in both species.

Overview of *Drosophila* Development

The life cycle of *Drosophila* is about ten days long with a number of distinct phases: the embryo, three larval stages, the pupal stage, and the adult stage (**Figure 23.3**). Internally, the cytoplasm of the unfertilized egg is organized into a series of maternally constructed molecular gradients that play a key role in determining the developmental fates of nuclei located in specific regions of the embryo.

Immediately after fertilization, the zygote nucleus undergoes a series of nuclear divisions without cytokinesis [**Figure 23.4(a)** and **(b)**]. The resulting cell, with multiple nuclei, is called a syncytium. At about the tenth division, nuclei move to the periphery of the egg, where the cytoplasm contains localized gradients of maternally derived mRNA transcripts and proteins [**Figure 23.4(c)**]. After several more divisions, the nuclei become enclosed in plasma membranes [**Figure 23.4(d)**], forming a cellular layer around the outside edge of the embryo. Interactions between the nuclei and the cytoplasmic components of these cells initiate and direct the pattern of embryonic gene expression.

Germ cells, which in the adult, are destined to undergo meiosis and produce gametes, form at the posterior pole of the embryo [Figure 23.4(c) and (d)]. Nuclei in other regions of the embryo normally form somatic cells. In experiments where nuclei from these regions are transplanted into the posterior cytoplasm, they will form germ cells and not somatic cells, confirming that the cytoplasm at the posterior pole of the embryo contains maternal components that direct nuclei to form germ cells.

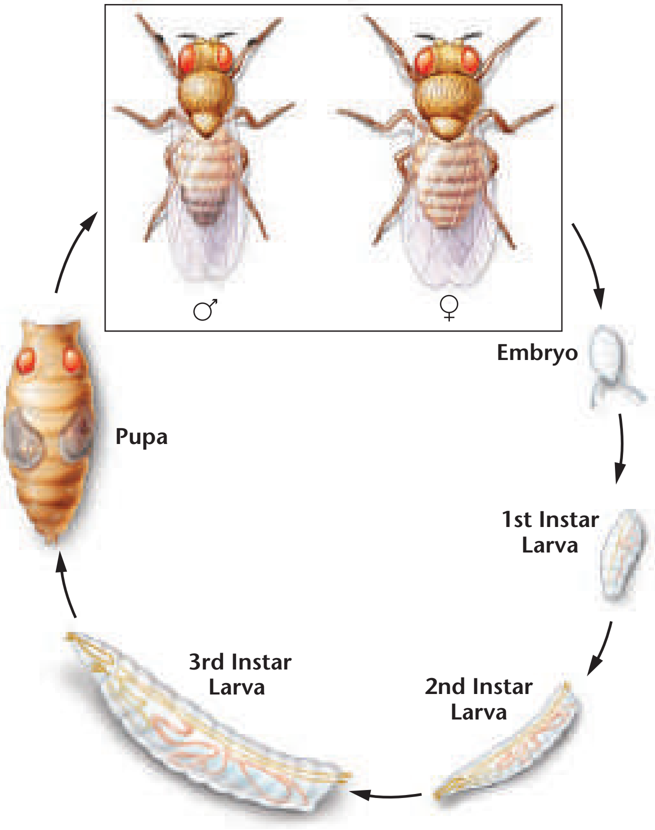
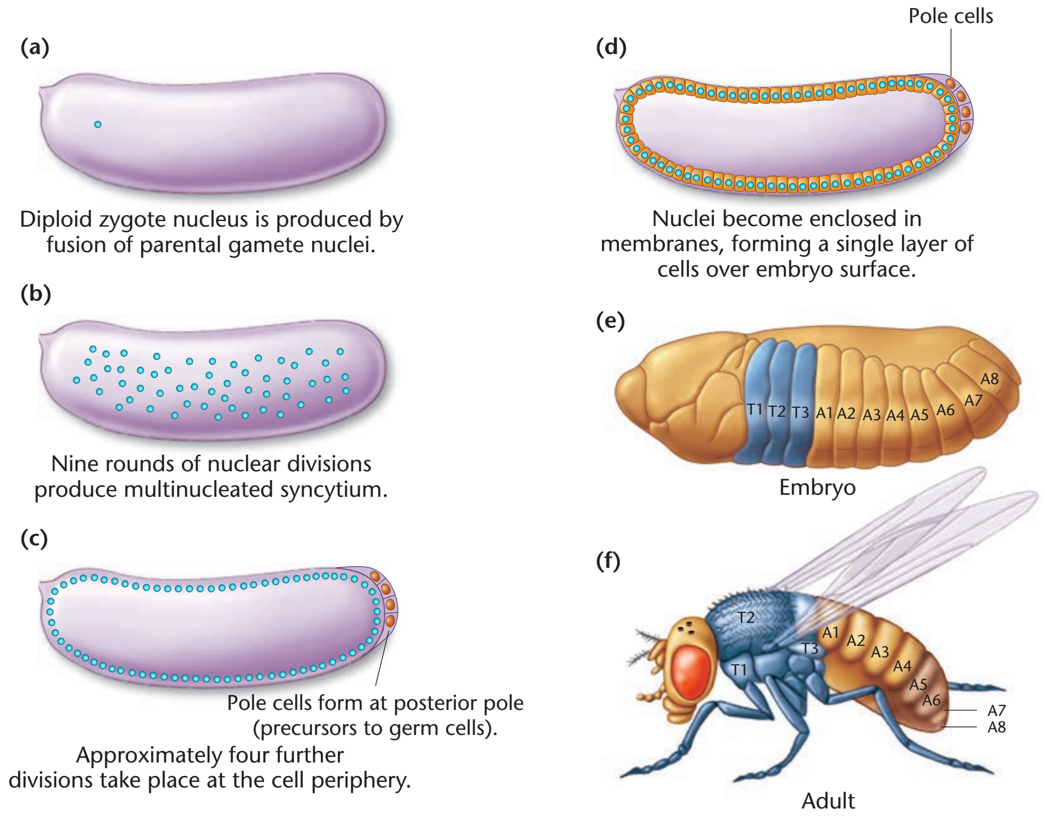


FIGURE 23.3 *Drosophila* life cycle.

Transcriptional programs activated by cytoplasmic components in somatic (non–germ-cell) nuclei form the embryo's anterior–posterior (front to back) and dorsal–ventral (upper to lower) axes of symmetry, leading to the formation of a segmented embryo [**Figure 23.4(e)**]. Under control of the *Hox* genes (discussed in section 23.5, these segments will give rise to the differentiated structures of the adult fly [**Figure 23.4(f)**].

Genetic Analysis of Embryogenesis

Two different gene sets control embryonic development in *Drosophila*: **maternal-effect genes** and **zygotic genes.** Products of maternal-effect genes (mRNA and/or proteins) are placed in the developing egg by the "mother" fly. Many of these products are distributed in a gradient or concentrated in specific regions of the egg cytoplasm. Female flies


(a)
(b)
(c)
Pole cells form at posterior pole (precursors to germ cells).
(d)
Pole cells
(e)
T1 T2 T3 A1 A2 A3 A4 A5 A6 A7 A8
Embryo
(f)
T1
T2
T3
A1
A2
A3
A4
A5
A6
A7
A8
Adult

Diploid zygote nucleus is produced by fusion of parental gamete nuclei.

Nine rounds of nuclear divisions produce multinucleated syncytium.

Approximately four further divisions take place at the cell periphery.

Nuclei become enclosed in membranes, forming a single layer of cells over embryo surface.

FIGURE 23.4 Early stages of embryonic development in *Drosophila*. (a) Fertilized egg with zygotic nucleus ($2n$), shortly after fertilization. (b) Nuclear divisions occur about every ten minutes. Nine rounds of division produce a multinucleate cell, the syncytial blastoderm. (c) At the tenth division, the nuclei migrate to the periphery or cortex of the egg, and four additional rounds of nuclear division occur. A small cluster of cells, the pole cells, form at the posterior pole about 2.5 hours after fertilization. These cells will form the germ cells of the adult. (d) About three hours after fertilization, the nuclei become enclosed in membranes, forming a single layer of cells over the embryo surface, creating the cellular blastoderm. (e) The embryo at about ten hours after fertilization. At this stage, the segmentation pattern of the body is clearly established. Behind the segments that will form the head, T1–T3 are thoracic segments, and A1–A8 are abdominal segments. (f) The adult fly showing the structures formed from each segment of the embryo.

homozygous for deleterious recessive mutations of maternal-effect genes are sterile. None of their embryos receive the wild-type maternal gene product encoded by the mutant allele, so all of the embryos develop abnormally and die. Maternal-effect genes encode transcription factors, receptors, and proteins that regulate gene expression. During development, these gene products activate or repress time- and location-specific programs of gene expression in the embryo.

Zygotic genes are those transcribed in the embryonic nuclei formed after fertilization. These products of the embryonic genome are differentially transcribed in specific regions of the embryo in response to the distribution of maternal-effect proteins. Deleterious recessive mutations in these genes can lead to embryonic lethality in homozygotes.

Much of our knowledge about the genes that regulate *Drosophila* development is based on the work of Edward Lewis, Christiane Nüsslein-Volhard, and Eric Wieschaus, who were awarded the 1995 Nobel Prize for Physiology or Medicine. Lewis initially identified and studied one of these regulatory genes. Nüsslein-Volhard and Wieschaus devised a strategy to identify all the genes that control segmentation in *Drosophila*. Their scheme required examining thousands of offspring of mutagenized flies, looking for recessive embryonic lethal mutations with defects in external structures. These mutations, called segmentation genes, were grouped into three classes: **gap**, **pair-rule,** and **segment polarity** genes.

On the basis of their observations, Nüsslein-Volhard and Wieschaus proposed a model that explains the formation of the body plan in *Drosophila*. In this model, the body plan is established by the hierarchical action of three classes of genes:

1. During egg formation, mRNA and proteins produced by transcription of the maternal genome (maternal-effect genes) are stored in the unfertilized egg. Many

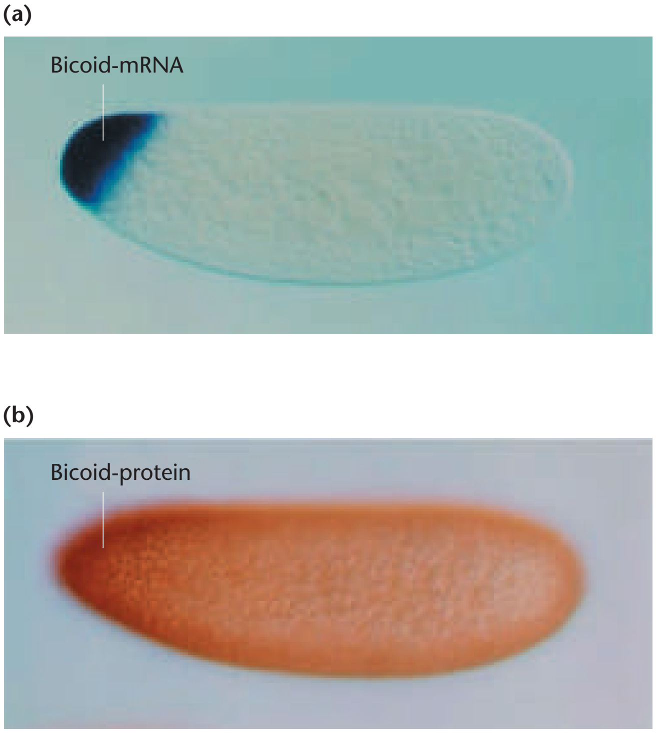


FIGURE 23.5 (a) Maternal mRNA transcripts of the *bicoid* gene are stored in the anterior tip of the *Drosophila* egg. (b) Translation of the bicoid mRNA produces a gradient of bicoid protein in cells at the anterior end of the embryo.

of these products are stored as positional gradients (**Figure 23.5**).

2. The positional information laid down by the maternal gene products is interpreted by the sequential transcription of zygotic **segmentation genes** (gap, pair-rule, and segment polarity genes) in cells of the embryo.
3. **Homeotic selector genes** specify the developmental fate of cells within each segment and determine which adult structures will be formed by each segment.

Their model is shown in **Figure 23.6**. Most maternal-effect gene products placed in the egg during oogenesis are activated immediately after fertilization and help establish the anterior–posterior axis of the embryo [**Figure 23.6(a)**]. These maternal genes encode transcription factors that activate the zygotic gap genes, whose expression divides the embryo into a series of regions corresponding to the head, thorax, and abdomen of the adult [**Figure 23.6(b)**]. Next, transcription factors encoded by gap genes activate the zygotic pair-rule genes, whose products divide the embryo into smaller regions about two segments wide [**Figure 23.6(c)**]. In turn, expression of the pair-rule genes activates the zygote's segment polarity genes, which divide each segment into anterior and posterior regions [**Figure 23.6(d)**].

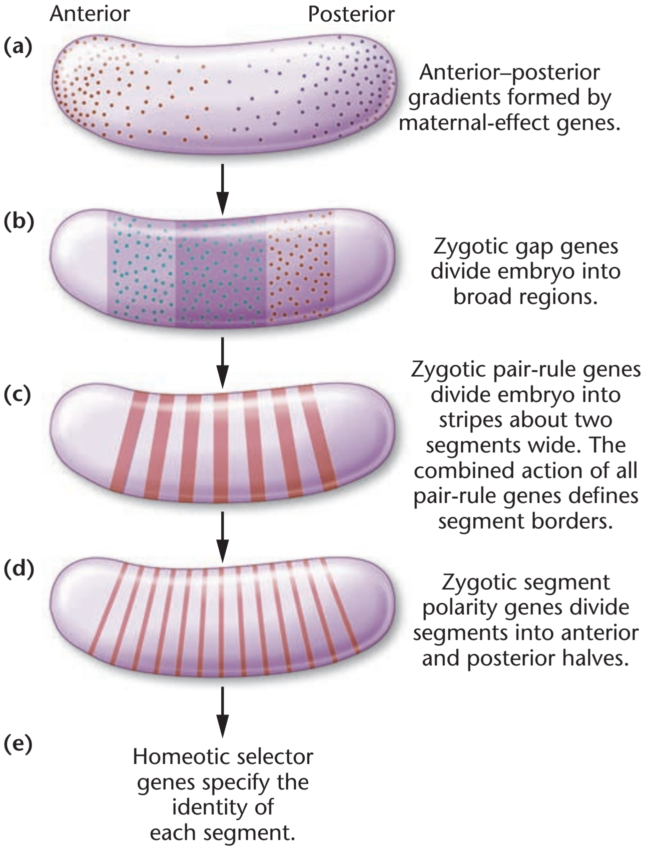

Anterior
Posterior
(a)
Anterior–posterior gradients formed by maternal-effect genes.
(b)
Zygotic gap genes divide embryo into broad regions.
(c)
Zygotic pair-rule genes divide embryo into stripes about two segments wide. The combined action of all pair-rule genes defines segment borders.
(d)
Zygotic segment polarity genes divide segments into anterior and posterior halves.
(e)
Homeotic selector genes specify the identity of each segment.

FIGURE 23.6 (a) Progressive restriction of cell fate during development in *Drosophila*. Gradients of maternal proteins are established along the anterior–posterior axis of the embryo. (b–d) Three groups of segmentation genes progressively define the body segments. (e) Individual segments are given identity by the homeotic genes.

The collective action of the maternal-effect genes and the zygotic segmentation genes define the anterior–posterior axis and the number, size, and polarity of each segment. In the next stage of differentiation, these segments are the field of action for the homeotic (*Hox*) selector genes [**Figure 23.6(e)**].

NOW SOLVE THIS

23.1 Suppose you initiate a screen for maternal-effect mutations in *Drosophila* affecting external structures of the embryo and you identify more than 100 mutations that affect these structures. From their screenings, other researchers concluded that there are only about 40 maternal-effect genes. How do you reconcile these different results?

■ **Hint:** *This problem involves an understanding of how mutants are identified when adult Drosophila were exposed to mutagens. The key to its solution is an understanding of the differences between genes and alleles (see Chapter 3).*

23.4 Segment Formation and Body Plans in *Drosophila* and Mammals

To summarize, certain genes in the zygote's genome are activated or repressed according to a positional gradient of maternal gene products. Expression of three sets of segmentation genes divides the embryo into a series of segments along its anterior–posterior axis. These segmentation genes are transcribed in normally developing embryos, and mutations of these genes have embryo-lethal phenotypes.

Over 20 segmentation genes (**Table 23.1**) have been identified, and they are classified on the basis of their mutant phenotypes: (1) mutations in gap genes delete a group of adjacent segments, causing gaps in the normal body plan of the embryo, (2) mutations in pair-rule genes affect every other segment and eliminate a specific part of each affected segment, and (3) mutations in segment polarity genes cause defects in portions of each segment.

In addition to these three sets of genes that determine the anterior–posterior axis of the developing embryo, another set of genes determines the dorsal–ventral axis of the embryo. Our discussion will be limited to the gene sets involved in the anterior–posterior axis. Let us now examine members of each group in greater detail.

Gap Genes

Transcription of **gap genes** in the embryo is controlled by maternal gene products laid down in gradients in the egg. Gap genes also cross-regulate each other to define the early stage of the body plan. Transcription of wild-type gap genes (which encode transcription factors) divides the embryo into a series of broad regions that become the head, thorax, and abdomen. Within these regions, different combinations of gene activity eventually specify both the type of segment that forms and the proper order of segments in the body of the larva, pupa, and adult. Mutant alleles of these genes produce large gaps in the embryo's segmentation pattern. *Hunchback* mutants lose head and thorax structures, *Krüppel* mutants lose thoracic and abdominal structures, and *knirps* mutants lose most abdominal structures. Expression domains of the gap genes in different parts of the embryo correlate roughly with the location of their mutant phenotypes: *hunchback* at the anterior, *Krüppel* in the middle (**Figure 23.7**), and *knirps* at the posterior. As mentioned earlier, gap genes encode transcription factors that control the expression of pair-rule genes.

TABLE 23.1 Segmentation Genes in *Drosophila*

Gap Genes	Pair-Rule Genes	Segment Polarity Genes
Krüppel	*hairy*	*engrailed*
knirps	*even-skipped*	*wingless*
hunchback	*runt*	*cubitis*
giant	*fushi-tarazu*	*hedgehog*
tailless	*paired*	*fused*
buckebein	*odd-paired*	*armadillo*
caudal	*odd-skipped*	*patched*
	sloppy-paired	*gooseberry*
		paired
		naked
		disheveled

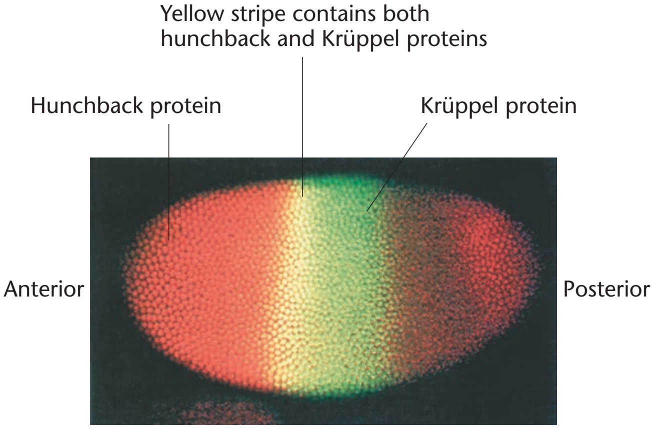


FIGURE 23.7 Visualization of gap gene expression in a *Drosophila* embryo. The hunchback protein is shown in orange, and Krüppel is indicated in green. The yellow stripe is created when cells contain both hunchback and Krüppel proteins. Each dot in the embryo is a nucleus.

Pair-Rule Genes

Pair-rule genes are expressed in a series of seven narrow bands or stripes of nuclei extending around the circumference of the embryo. The expression of this gene set does two things: first it establishes the boundaries of segments, and then it programs the developmental fate of the cells within each segment by controlling expression of the segment polarity genes. Mutations in pair-rule genes eliminate segment-size sections of the embryo at every other segment. At least eight pair-rule genes act to divide the embryo into a series of stripes. The transcription of the pair-rule genes is mediated by the action of maternal gene products and gap gene products. Initially, the boundaries of these stripes overlap, so that in each area of overlap, cells express a different combination of pair-rule genes (**Figure 23.8**). The resolution of boundaries in this segmentation pattern results from the combined activity of different transcription factors in adjacent segments (**Figure 23.9**).

Segment Polarity Genes

Expression of **segment polarity genes** is controlled by transcription factors encoded by pair-rule genes. Within each of the segments created by pair-rule genes, segment polarity

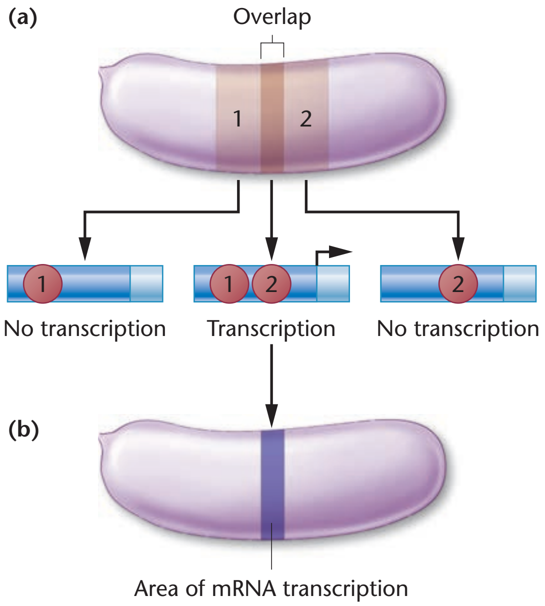


FIGURE 23.8 New patterns of gene expression can be generated by overlapping regions containing two different gene products. (a) Transcription factors 1 and 2 are present in an overlapping region of expression. If both transcription factors must bind to the promoter of a target gene to trigger expression, the gene will be active only in cells containing both factors (most likely in the zone of overlap). (b) The expression of the target gene in the restricted region of the embryo.

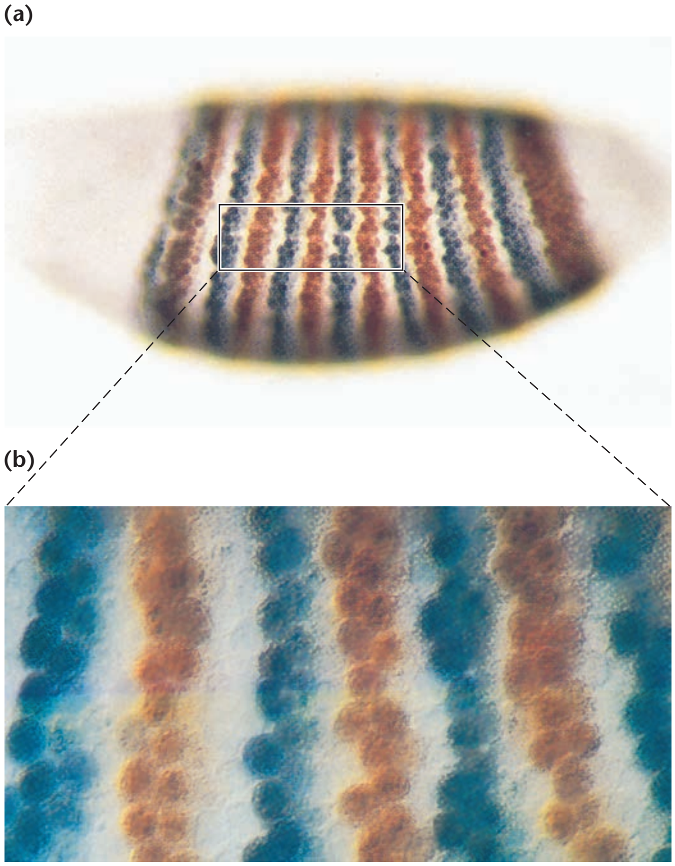


FIGURE 23.9 Stripe pattern of pair-rule gene expression in a *Drosophila* embryo. The nuclei of this embryo are stained to show patterns of expression of the genes *even-skipped* and *fushi-tarazu;* (a) low-power view and (b) inset: high-power view of the same embryo.

genes become active in a single band of cells that extends around the embryo's circumference (Figure 23.10). This divides the embryo into 14 segments. The products of the segment polarity genes control the cellular identity within each of them and establish the anterior–posterior pattern (the polarity) within each segment.

Segmentation Genes in Mice and Humans

We have seen that segment formation in *Drosophila* depends on the action of three sets of segmentation genes. Are these genes found in humans and other mammals, and do they control aspects of embryonic development in these organisms? To answer these questions, let's examine *runt,* one of the pair-rule genes in *Drosophila.* Later in development, it controls aspects of sex determination and formation of the nervous system. The gene encodes a protein that regulates transcription of its target genes. Runt contains a 128-amino-acid DNA-binding region (called the runt domain) that is highly conserved in *Drosophila*, mouse, and human proteins. In fact, *in vitro* experiments show that the *Drosophila* and mouse runt proteins are functionally interchangeable. In mice, *runt* is expressed early in development and controls formation of blood cells, bone, and the genital system. Although the target gene sets controlled by *runt* are different in *Drosophila* and the mouse, in both organisms, expression of *runt* specifies the fate of uncommitted cells in the embryo by regulating transcription of target genes.

In humans, mutations in *RUNX2,* a human homolog of *runt,* causes cleidocranial dysplasia (CCD), an autosomal dominantly inherited trait. Those affected with CCD have a hole in the top of their skull because bone does not form in the membranous gap known as the fontanel. Their collar bones (clavicles) do not develop, enabling affected

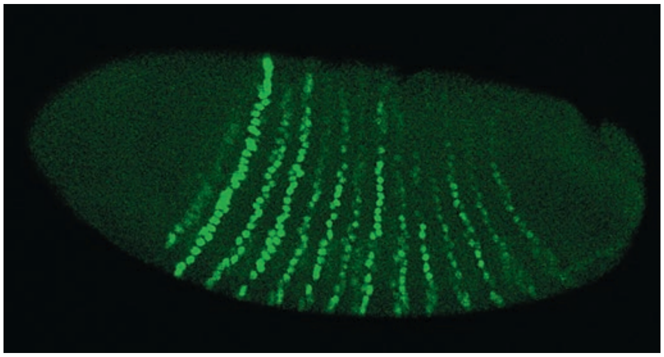
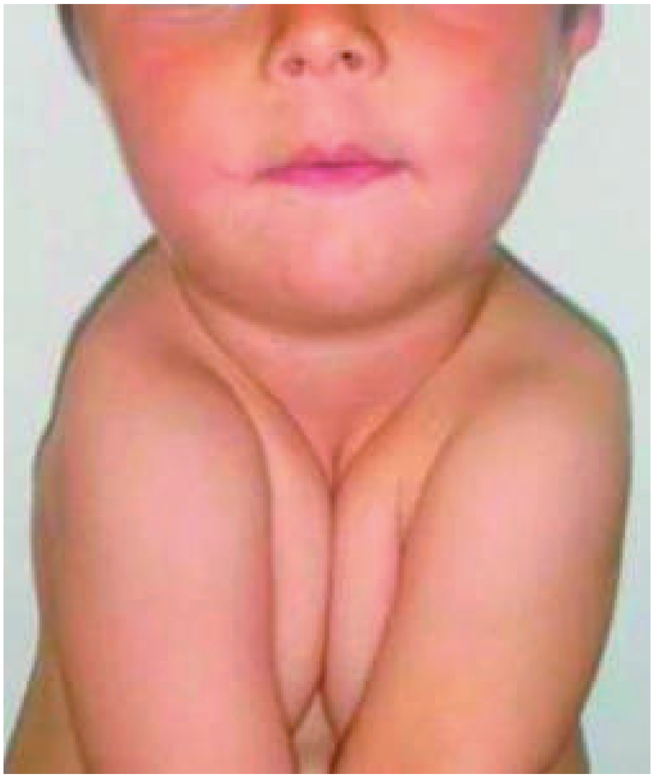


FIGURE 23.10 The 14 stripes of expression of the segment polarity gene *engrailed* in a *Drosophila* embryo.



individuals to fold their shoulders across their chest (**Figure 23.11**). Mice with one mutant copy of the *runt* homolog have a phenotype similar to that seen in humans; mice with two mutant copies of the gene have no bones at all. Their skeletons contain only cartilage (**Figure 23.12**), much like sharks, emphasizing the role of *runt* as an important gene controlling the initiation of bone formation in both mice and humans.

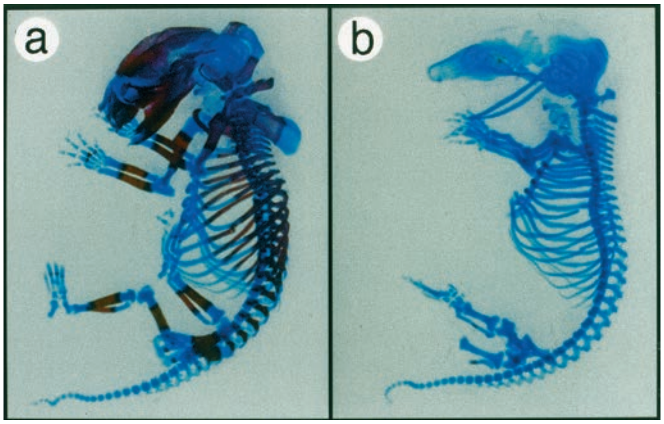


FIGURE 23.12 Bone formation in normal mice and mutants for the *runt* gene *Cbfa1*. (a) Normal mouse embryos at day 17.5 show cartilage (blue) and bone (brown). (b) The skeleton of a 17.5-day homozygous mutant embryo. Only cartilage has formed in the skeleton. There is complete absence of bone formation in the mutant mouse. Expression of a normal copy of the *Cbfa1* gene is essential for specifying the developmental fate of bone-forming osteoblasts.

FIGURE 23.11 A boy affected with cleidocranial dysplasia (CCD). This disorder, inherited as an autosomal dominant trait, is caused by mutation in a human *runt* gene, *RUNX2*. Affected heterozygotes have a number of skeletal defects, including a hole in the top of the skull where the infant fontanel fails to close, and collar bones that do not develop, or form only small stumps. Because the collar bones do not form, individuals with CCD can fold their shoulders across their chests.

The *runt* domain sequence similarity in *Drosophila*, mice, and humans and the ability of the mouse *runt* gene to replace the *Drosophila* version in fly development all indicate that the same segmentation genes are found in organisms separated from a common ancestor by millions of years.

23.5 Homeotic Selector Genes Specify Body Parts of the Adult

As segment boundaries are established by expression of segmentation genes, the homeotic (from the Greek word for "same") genes are activated. Expression of homeotic selector genes determines which adult structures will be formed by each body segment. They are called selector genes because action of these genes selects one developmental pathway appropriate for a given segment from several alternatives. In *Drosophila,* these pathways lead to the formation of the antennae, mouthparts, legs, wings, thorax, and abdomen. Mutants of these genes are called **homeotic mutants** because one segment is transformed so that it forms the same structure as another segment. For example, the wild-type allele of *Antennapedia* (*Antp*) specifies the developmental pathway leading to formation of a leg on the second segment of the thorax. Dominant gain-of-function *Antp* mutations cause this gene to be expressed in the head *and* the thorax. The result is that mutant flies have a leg on their head instead of an antenna (**Figure 23.13**).

Hox Genes in *Drosophila*

The *Drosophila* genome contains two clusters of homeotic selector genes (called *Hox* genes) on chromosome 3 that encode transcription factors (**Table 23.2**). The *Antennapedia* (*ANT-C*) cluster contains five genes that specify

TABLE 23.2 *Hox* Genes of *Drosophila*

Antennapedia Complex	*Bithorax* Complex
labial	*Ultrabithorax*
Antennapedia	*abdominal A*
Sex combs reduced	*Abdominal B*
Deformed	
proboscipedia	

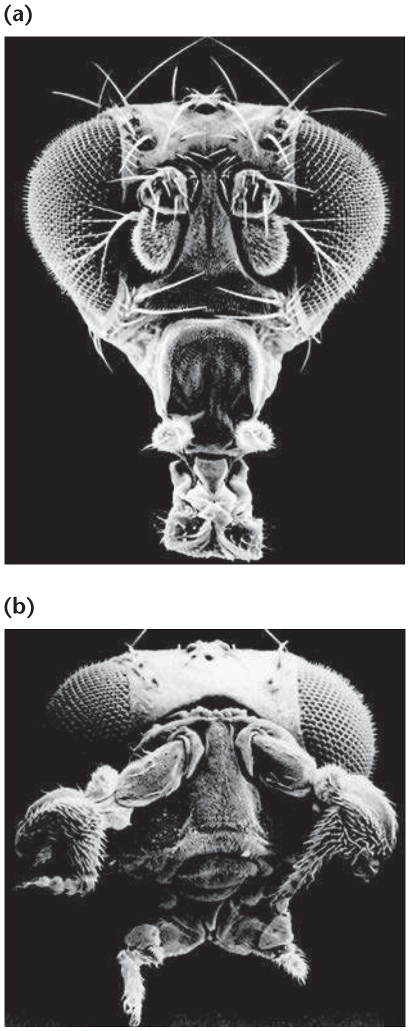


FIGURE 23.13 *Antennapedia* (*Antp*) mutation in *Drosophila*. (a) Head from wild-type *Drosophila*, showing the antenna and other head parts. (b) Head from an *Antp* mutant, showing the replacement of normal antenna structures with legs. This is caused by activation of the *Antp* gene in the head region.

structures in the head and the first two segments of the thorax [Figure 23.14(a)]. The second cluster, the *bithorax* (*BX-C*) complex, contains three genes that specify structures in the posterior portion of the second thoracic segment, the entire third thoracic segment, and the abdominal segments [Figure 23.14(b)].

Hox genes (listed in Table 23.2) from a wide range of species have two properties in common. First, each contains a highly conserved 180-bp nucleotide sequence known as a **homeobox**. (*Hox* is a contraction of homeobox.) The homeobox encodes a DNA-binding region of 60 amino acids known as a **homeodomain**. Second, in most species, expression of *Hox* genes is colinear with the anterior to posterior organization of the body. In other words, genes at one end of the cluster (the 3′ end) are expressed at the anterior end of the embryo, those in the middle are expressed in the middle of the embryo, and genes at the other end of a cluster (the 5′ end) are expressed at the embryo's posterior region (Figure 23.15). Although first identified in *Drosophila, Hox* genes are found in the genomes of most eukaryotes with segmented body plans, including zebrafish, frogs, mice, and humans (Figure 23.16).

To summarize, genes that control development in *Drosophila* act in a temporally and spatially ordered cascade, beginning with the genes that establish the anterior–posterior (and dorsal–ventral) axis of the early embryo. Gradients of maternal mRNAs and proteins along the anterior–posterior axis activate gap genes, which subdivide the embryo into broad bands. Gap genes in turn activate pair-rule genes,

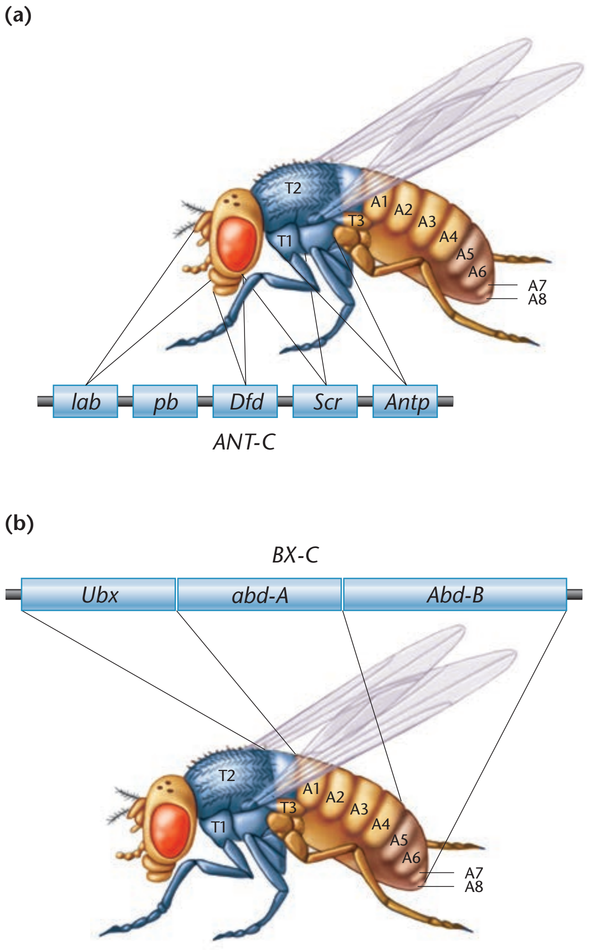
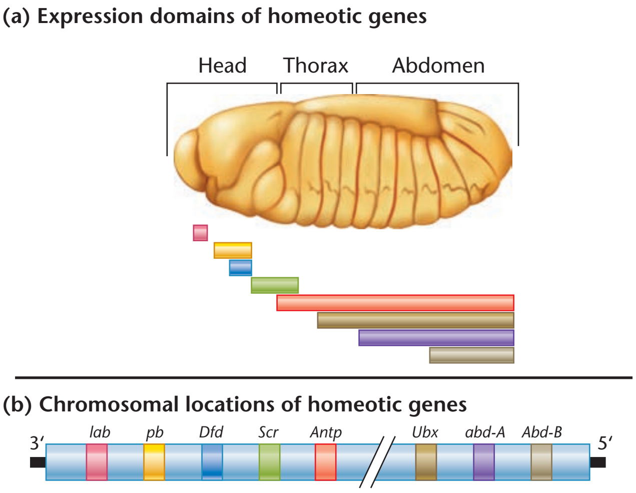


FIGURE 23.14 Genes of the *Antennapedia* complex and the adult structures they specify. (a) In the *ANT-C* complex, the *labial* (*lab*) and *Deformed* (*Dfd*) genes control the formation of head segments. The *Sex comb reduced* (*Scr*) and *Antennapedia* (*Antp*) genes specify the identity of the first two thoracic segments, T1 and T2. The remaining gene in the complex, *proboscipedia* (*pb*), may not act during embryogenesis but may be required to maintain the differentiated state in adults. In mutants, the labial palps are transformed into legs. (b) In the *BX-C* complex, *Ultrabithorax* (*Ubx*) controls formation of structures in the posterior compartment of T2 and structures in T3. The two other genes, *abdominal A* (*abdA*) and *Abdominal B* (*AbdB*), specify the segmental identities of the eight abdominal segments (A1–A8).


(a) Expression domains of homeotic genes

(b) Chromosomal locations of homeotic genes

FIGURE 23.15 The colinear relationship between the spatial pattern of expression and chromosomal locations of homeotic genes in *Drosophila.* (a) *Drosophila* embryo and the domains of homeotic gene expression in the embryonic epidermis and central nervous system. (b) Chromosomal location of homeotic selector genes. Note that the order of genes on the chromosome correlates with the sequential anterior borders of their expression domains.

which divide the embryo into segments. The final group of segmentation genes, the segment polarity genes, divides each segment into anterior and posterior regions arranged linearly along the anterior–posterior axis. The segments are then given identity by action of the *Hox* genes. Therefore, this progressive restriction of developmental potential of the *Drosophila* embryo's cells (all of which occurs during the first third of embryogenesis) involves a cascade of gene action, with regulatory proteins acting on transcription, translation, and signal transduction.

Hox Genes and Human Genetic Disorders

Although first described in *Drosophila*, *Hox* genes are found in the genomes of all animals, where they play a fundamental role in shaping the body and its appendages. In vertebrates, the conservation of sequence, the order of genes in the *Hox* clusters, and their pattern of expression suggests that, as in *Drosophila*, these genes control development along the anterior–posterior axis and the formation of appendages. However, in vertebrates, there are four clusters of *Hox* genes—*HOXA*, *HOXB*, *HOXC*, and *HOXD* due to cluster duplication during vertebrate evolution—instead of a single cluster as in *Drosophila*. This means that in vertebrates, not just one, but a combination of two to four *Hox* genes is involved in forming specific structures. As a result, in vertebrates, mutations in individual *Hox* genes do not typically produce complete transformation as in *Drosophila*, where mutation of a single *Hox* gene can transform a haltere into a wing (see the chapter opening photo). The role of *HOXD* genes in human development was confirmed by the discovery that several inherited limb malformations are caused by mutations in *HOXD* genes. For example, mutations in *HOXD13* cause synpolydactyly (SPD), a malformation characterized by extra fingers and toes, and abnormalities in bones of the hands and feet.

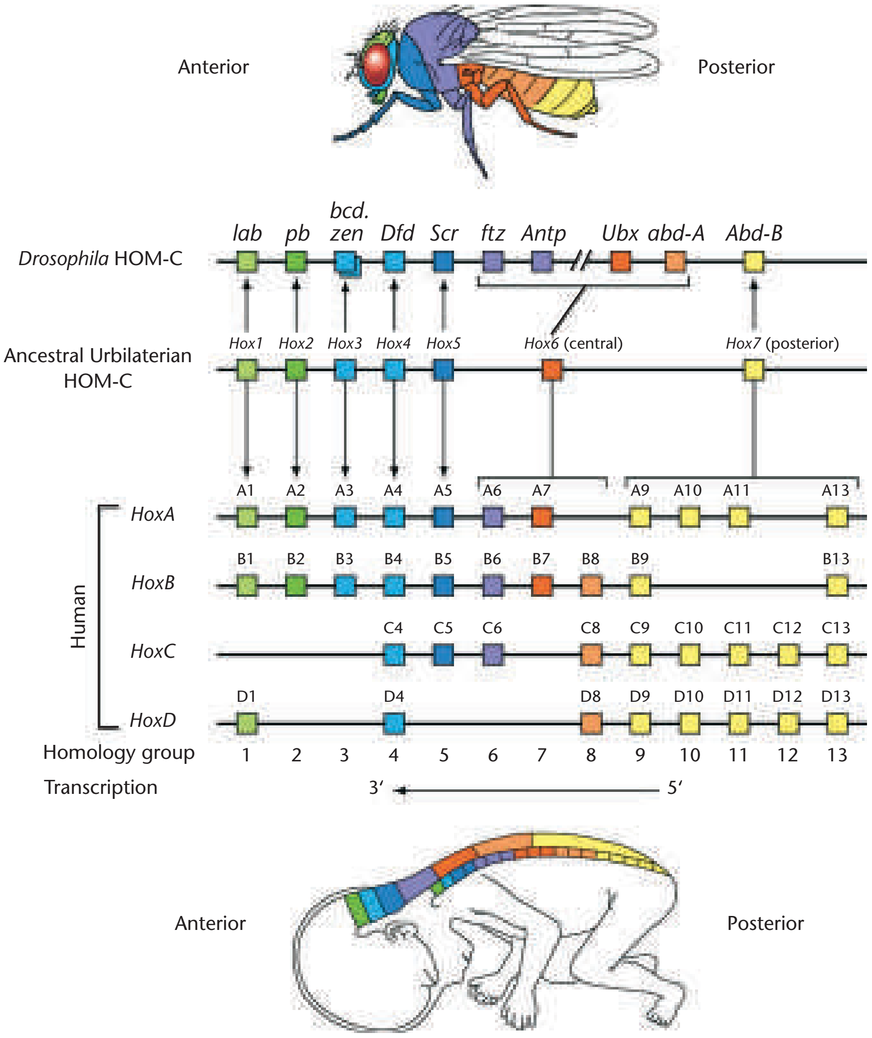


FIGURE 23.16 Conservation of organization and patterns of expression in *Hox* genes. (Top) The structures formed in adult *Drosophila* are shown, with the colors corresponding to members of the *Hox* cluster that control their formation. (Bottom) The arrangement of the *Hox* genes in an early human embryo. As in *Drosophila,* genes at the (3′ end) of the cluster form anterior structures, and genes at the (5′ end) of the cluster form posterior structures.

NOW SOLVE THIS

23.2 In *Drosophila*, both *fushi tarazu* (*ftz*) and *engrailed* (*eng*) genes encode homeobox transcription factors and are capable of eliciting the expression of other genes. Both genes work at about the same time during development and in the same region to specify cell fate in body segments. To discover if *ftz* regulates the expression of *engrailed*, if *engrailed* regulates *ftz*, or if both are regulated by another gene, you perform a mutant analysis. In *ftz* embryos (*ftz*/*ftz*) engrailed protein is absent; in *engrailed* embryos (*eng*/*eng*) *ftz* expression is normal. What does this tell you about the regulation of these two genes—does the *engrailed* gene regulate *ftz*, or does the *ftz* gene regulate *engrailed*?

- **Hint:** *This problem involves an understanding of how genes are regulated at different stages of preadult development in* Drosophila. *The key to its solution lies in using the results of the mutant analysis to determine the timing of expression of the two genes being examined.*

For more practice, see Problems 15, 16, and 18.

23.6 Plants Have Evolved Developmental Regulatory Systems That Parallel Those of Animals

Plants and animals diverged from a common ancestor about 1.6 billion years ago, after the origin of eukaryotes and probably before the rise of multicellular organisms. Genetic analysis of mutants and genome sequencing in plants and animals indicate that basic mechanisms of developmental pattern formation evolved independently in animals and plants. We have already examined the genetic systems that control development and pattern formation in animals, using *Drosophila* as a model organism, and will now briefly examine these systems in plants.

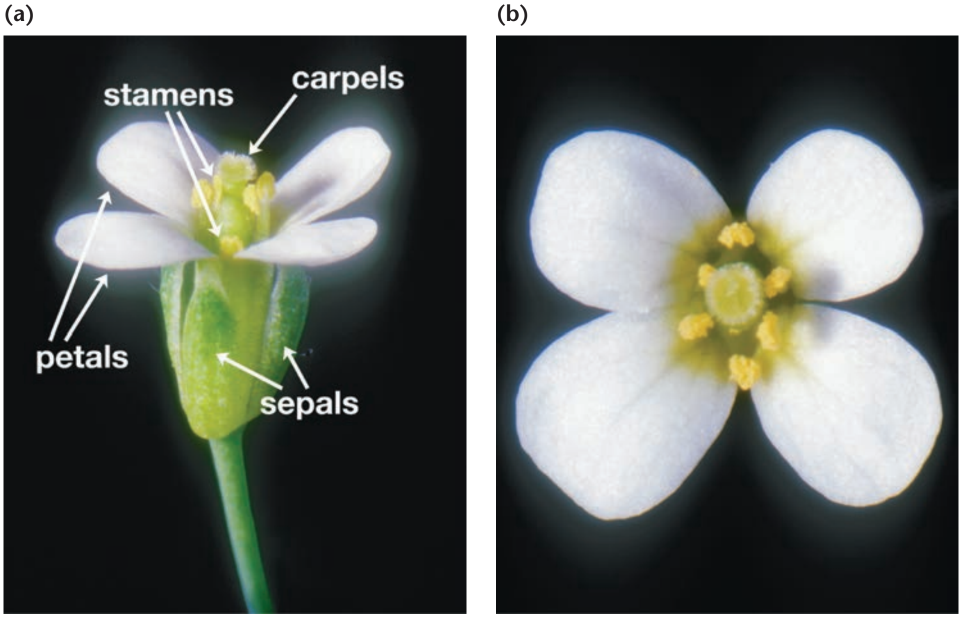


FIGURE 23.17 (a) Parts of the *Arabidopsis* flower. The floral organs are arranged concentrically. The sepals form the outermost ring, followed by petals and stamens, with carpels on the inside. (b) View of the flower from above.

TABLE 23.3 Homeotic Selector Genes in *Arabidopsis**

Class A	*APETALA1 (AP1)*
	APETALA2 (AP2)
Class B	*APETALA3 (AP3)*
	PISTILLATA (P1)
Class C	*AGAMOUS (AG)*

*By convention, wild-type genes in *Arabidopsis* use capital letters.

In plants, pattern formation has been extensively studied using flower development in *Arabidopsis thaliana*, a small plant in the mustard family, as a model organism. A cluster of undifferentiated cells, called the *floral meristem*, gives rise to flowers (**Figure 23.17**). Each flower consists of four organs—sepals, petals, stamens, and carpels—that develop from concentric rings of cells within the meristem [**Figure 23.18(a)**]. Each organ develops from a different concentric ring, or whorl of cells.

Homeotic Genes in *Arabidopsis*

Three classes of floral homeotic genes control the development of these organs (**Table 23.3**). Acting alone, class A genes specify sepals; class A and class B genes expressed together specify petals. Acting together, class B and class C genes control stamen formation. Class C genes acting alone specify carpels. During flower development [**Figure 23.18(b)**], class A genes are active in whorls 1 and 2 (sepals and petals), class B genes are expressed in whorls 2 and 3 (petals and stamens), and class C genes are expressed in whorls 3 and 4 (stamens and carpels). The organ formed depends on the expression pattern of the three gene classes. In whorl 1, expression of class A genes alone causes sepals to form. Expression of class A *and* class B genes in whorl 2 leads

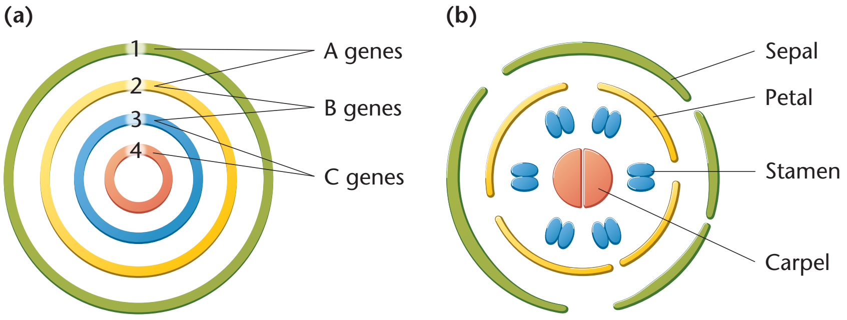


FIGURE 23.18 Cell arrangement in the floral meristem. (a) The four concentric rings, or whorls, labeled 1–4, give rise to (b) arrangement of the sepals, petals, stamens, and carpels, respectively, in the mature flower.

to petal formation. Expression of class B and class C genes in whorl 3 leads to stamen formation. In whorl 4, expression of class C genes alone causes carpel formation.

As in *Drosophila*, mutations in homeotic genes cause organs to form in abnormal locations. For example, in *APETALA2* mutants (*AP2*), the order of organs is carpel, stamen, stamen, and carpel instead of the normal order, sepal, petal, stamen, and carpel [**Figure 23.19(a)** and **(b)**]. In class B loss-of-function mutants (*AP3*, *P1*), petals become sepals, and stamens are transformed into carpels [**Figure 23.19(c)**], and the order of organs becomes sepal, sepal, carpel, carpel. Plants carrying a mutation for the class C gene *AGAMOUS* will have petals in whorl 3 (instead of stamens) and sepals in whorl 4 (instead of carpels), and the order of organs will be sepal, petal, petal, and sepal [**Figure 23.19(d)**].

Divergence in Homeotic Genes

Drosophila and *Arabidopsis* use different sets of nonhomologous master regulatory genes to establish the body axis and specify the identity of structures along the axis. In *Drosophila*, this task is accomplished in part by the *Hox* genes, which encode a set of transcription factors sharing a homeobox domain. In *Arabidopsis*, the floral homeotic genes belong to a different family of transcription factors, called the **MADS-box proteins**, characterized by a common sequence of 58 amino acids with no similarity in amino acid sequence or protein structure with the *Hox* genes. Both gene sets encode transcription factors, both sets are master regulators of development expressed in a pattern of overlapping domains, and both specify identity of structures.

Reflecting their evolutionary origin from a common ancestor, the genomes of both *Drosophila* and *Arabidopsis* contain members of the homeobox and MADS-box genes, but these genes have been adapted for different uses in the plant and animal kingdoms. This indicates that developmental mechanisms evolved independently in each group.

In both plants and animals, the action of transcription factors depends on changes in chromatin structure that make genes available for expression. Mechanisms of transcription initiation are conserved in plants and animals, as is reflected in the homology of genes in *Drosophila* and *Arabidopsis* that maintain patterns of expression initiated by regulatory gene sets. Action of the floral homeotic genes is controlled by a gene called *CURLY LEAF*. This gene shares significant homology with members of a *Drosophila* gene family called *Polycomb*. This family of regulatory genes controls expression of homeobox genes during development. Both *CURLY LEAF* and *Polycomb* encode proteins that

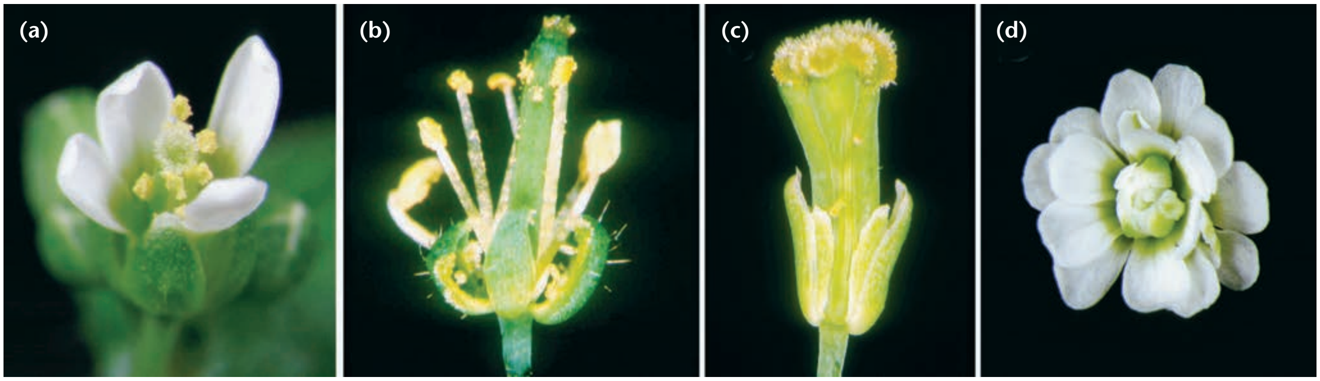


FIGURE 23.19 (a) Wild-type flowers of *Arabidopsis* have (from outside to inside) sepals, petals, stamens, and carpels. (b) A homeotic *APETALA2* mutant flower has carpels, stamens, stamens, and carpels. (c) *PISTILLATA* mutants have sepals, sepals, carpels, and carpels. (d) *AGAMOUS* mutants have petals and sepals at places where stamens and carpels should form.

alter chromatin conformation and shut off gene expression. Thus, although different genes are used to control development, both plants and animals use an evolutionarily conserved mechanism to regulate expression of these gene sets.

23.7 *C. elegans* Serves as a Model for Cell–Cell Interactions in Development

During development in multicellular organisms, cell–cell interactions influence the transcriptional programs and developmental fate of the interacting cells and surrounding cells. Cell–cell interaction is an important process in the embryonic development of most eukaryotic organisms, including *Drosophila*, mice, and humans.

Signaling Pathways in Development

In early development, animals use a number of signaling pathways to regulate development; after organ formation begins, additional pathways are added to those already in use. These newly activated pathways act both independently and in coordinated networks to generate specific transcriptional patterns. Signal networks establish anterior–posterior polarity and body axes, coordinate pattern formation, and direct the differentiation of tissues and organs. The signaling pathways used in early development and some of the developmental processes they control are listed in **Table 23.4**. The Wnt/β-catenin signaling system arose early in animal evolution. It is complex, highly conserved, and incorporates multibranched pathways involved in basic aspects of animal development. Many parts of this system are being explored by researchers in the hope that understanding how Wnt/β-catenin signaling controls development will result in treatments for disorders associated with pathway malfunctions.

Here, we will focus on an introduction to the components and interactions of another signaling system—the **Notch signaling pathway**—and examine its role in the development of the vulva in the nematode *Caenorhabditis elegans*.

TABLE 23.4 Signaling Pathways Used in Early Embryonic Development

Wnt Pathway
Dorsalization of body
Female reproductive development
Dorsal–ventral differences
TGF-β Pathway
Mesoderm induction
Left–right asymmetry
Bone development
Hedgehog Pathway
Notochord induction
Somitogenesis
Gut/visceral mesoderm
Receptor Tyrosine Kinase Pathway
Mesoderm maintenance
Notch Signaling Pathway
Blood cell development
Neurogenesis
Retina development

The Notch Signaling Pathway

The genes in the Notch pathway are named after the *Drosophila* mutants that were used to identify components of this signal transduction system (*Notch* mutants have an indentation or notch in their wings). The Notch signal system works through direct cell–cell contact to control the developmental fate of interacting cells. The *Notch* gene (and the equivalent gene in other organisms) encodes a signal receptor protein embedded in the plasma membrane (**Figure 23.20**). The signal is another membrane protein encoded by the *Delta* gene (and its equivalents). Because

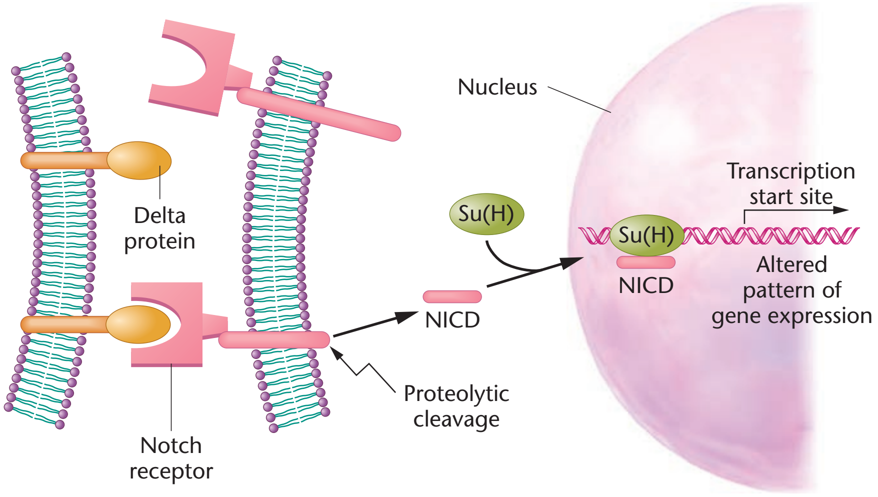
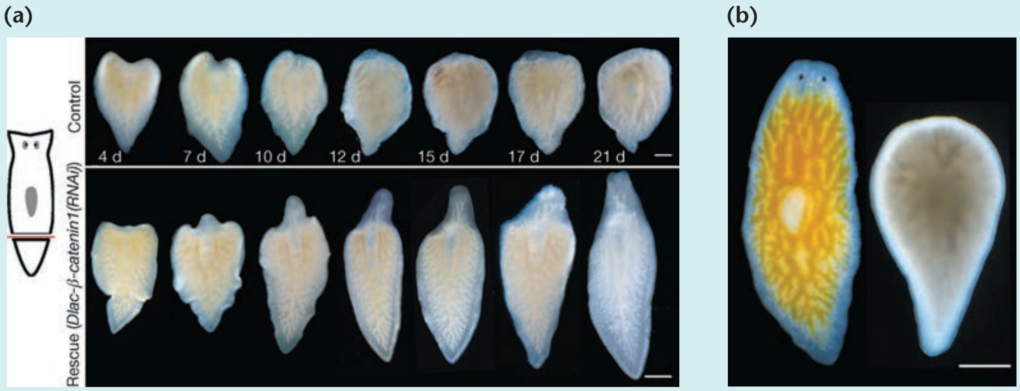


FIGURE 23.20 Components of the Notch signaling pathway in *Drosophila*. The cell carrying the Delta transmembrane protein is the sending cell; the cell carrying the transmembrane Notch protein receives the signal. Binding of Delta to Notch triggers a proteolytic-mediated activation of transcription. The fragment cleaved from the cytoplasmic side of the Notch protein, called the Notch intracellular domain (NICD), combines with the Su(H) protein and moves to the nucleus where it activates a program of gene transcription.

MODERN APPROACHES TO UNDERSTANDING GENE FUNCTION

Downregulating a Single Gene Reveals Secrets to Head Regeneration in Planaria



(a) A timed photo series showing the effect of control (top) or β-*catenin*1 *RNAi treated tail pieces (bottom). Control pieces fail to regenerate heads; 16 of 24 tail pieces treated with RNAi formed heads. (b)* β-*catenin*1 *RNAi treated tail pieces (yellow) but not control pieces (gray) developed heads that could ingest phenol-red labeled food particles, demonstrating functionality of the regenerated heads.*

The regenerative properties of some organisms are truly remarkable. Animals such as hydra, planaria, and salamanders can regenerate organs and entire body parts after they have been damaged or even amputated. Developmental biologists and other scientists have been very interested in understanding why these organisms possess regenerative properties. What are the evolutionary mechanisms that have ultimately determined why some organisms can regenerate and others cannot? This information is also relevant when applied to humans for the benefits of tissue and organ repair and replacement.

In 2013, *Nature* published papers from three teams identifying a signaling pathway for head regeneration from stem cells *in Planaria*. In many species of *Planaria*, animals can regenerate a head from a tail piece. However, this work focused on planarian species that are regeneration deficient and are unable to regenerate a head from a tail fragment. The logic for using these species is that if scientists can determine why such planarians cannot regenerate heads, it will likely provide important clues about the genetic pathways necessary for regeneration in flatworms and other animals.

In this chapter you were introduced to the Wnt/β-catenin pathway and its roles in early embryonic development in multicellular organisms, including humans. Wnt protein causes an accumulation of β-catenin, which regulates expression of genes important for development and tissue repair and maintenance. Mutations of the Wnt/β-catenin pathway are associated with developmental defects and other phenotypes such as tumor formation. Previous studies on regeneration-capable planarians have shown that enhanced activation of the Wnt/β-catenin pathway results in tail growth from anterior wounds, producing animals with two tails. Conversely, in posterior wounds reduced activity of the pathway results in head growth, instead of a tail, creating animals with two heads.

One hypothesis to explain the inability of planaria to regenerate a head from tail fragments is that specific levels of Wnt/β-catenin pathway activation are required for head regeneration. To test this hypothesis we highlight a particular experiment that applied **RNA interference (RNAi)**. Recall from earlier in the text (see Chapter 17), that RNAi is a sequence-specific method for silencing RNA molecules. Using the regeneration-deficient planaria species *Dugesia lacteum* (*D. lacteum*), researchers tried a head "rescue" experiment (see figure panel a). They removed tail pieces of *D. lacteum* and injected these segments with double-stranded RNA (dsRNA) designed to silence β-*catenin*1 mRNA by RNAi.

Results:

β-*catenin*1 RNAi treatment rescued head regeneration in 16 out of 24 tail pieces (figure panel a). Researchers analyzed marker genes associated with head and central nervous system structures to assess head structure. They tested functionality of rescued heads by demonstrating that β-*catenin*1 RNAi regenerated planaria could ingest dye-labeled food particles.

In separate experiments, the authors noted that a failure to upregulate expression of a Wnt inhibitor gene called *notum* in *D. lacteum* tail pieces may possibly explain failed head regeneration.

Conclusions:

Researchers concluded that in regeneration-deficient species such as *D. lacteum* there is insufficient downregulation (inhibition) of the Wnt/β-catenin to allow heads to regenerate. Thus downregulating this pathway by RNAi leads to head regeneration. These results suggest that regeneration-capable species can downregulate the Wnt/β-catenin when necessary. Most impressive is the finding that a single signaling pathway is sufficient to induce regeneration of functioning heads. Understanding which genes regulate the Wnt/β-catenin pathway in both regeneration-capable and regeneration-deficient species is an important next step in this research. There is every reason to expect that learning about molecular mechanisms involved in regeneration in planaria will help scientists improve regenerative capabilities in humans.

(continued)

Modern Approaches to Understanding Gene Function, continued

References:

Liu, S-Y., et al. (2013). Reactivating head regrowth in a regeneration-deficient planarian species. *Nature* 500:81–84.

Sikes, J. M. and Newmark P. A. (2013). Restoration of anterior regeneration in a planarian with limited regenerative ability. *Nature* 500:77–80.

Umesono, Y., et al. (2013). The molecular logic for planarian regeneration along the anterior-posterior axis. *Nature* 500:73–76.

Questions to Consider:

1. It is thought that high levels of the protein ERK are required for head formation in *D. lacteum* and that the Wnt/β-catenin pathway may suppress ERK. Propose an experiment to test this hypothesis. Based on the experimental approach described here, what experiments could you do to determine if ERK is involved in head regeneration?
2. Results of these experiments suggests that regenerating complex body parts may not always require a detailed understanding of many different individual signaling pathways. Defend or refute this statement keeping in mind potential species-specific applications (not just human tissue repair) of repairing regeneration defects.

both the signal and receptor are membrane proteins, the Notch signal system works only between adjacent cells. When the Delta protein from one cell binds to the Notch receptor protein on a neighboring cell, the cytoplasmic tail of the Notch protein is cleaved off and binds to a cytoplasmic protein encoded by the *Su(H)* (suppressor of *Hairless*) gene. This protein complex moves into the nucleus and binds to transcriptional cofactors, activating transcription of a gene set that controls a specific developmental pathway.

One of the main roles of the Notch signal system is to specify different developmental fates for equivalent cells in a population. In its simplest form, this interaction involves two neighboring cells that are developmentally equivalent. We will explore the role of the Notch signaling system in development of the vulva in *C. elegans*, after a brief introduction to nematode embryogenesis.

Overview of *C. elegans* Development

The nematode *C. elegans* is widely used to study the genetic control of development. There are several advantages in using this organism: (1) its genetics are well known, (2) its genome has been sequenced, and (3) adults contain a small number of cells that follow a highly deterministic developmental program. Adult nematodes are about 1 mm long and develop from a fertilized egg in about two days (**Figure 23.21**). The

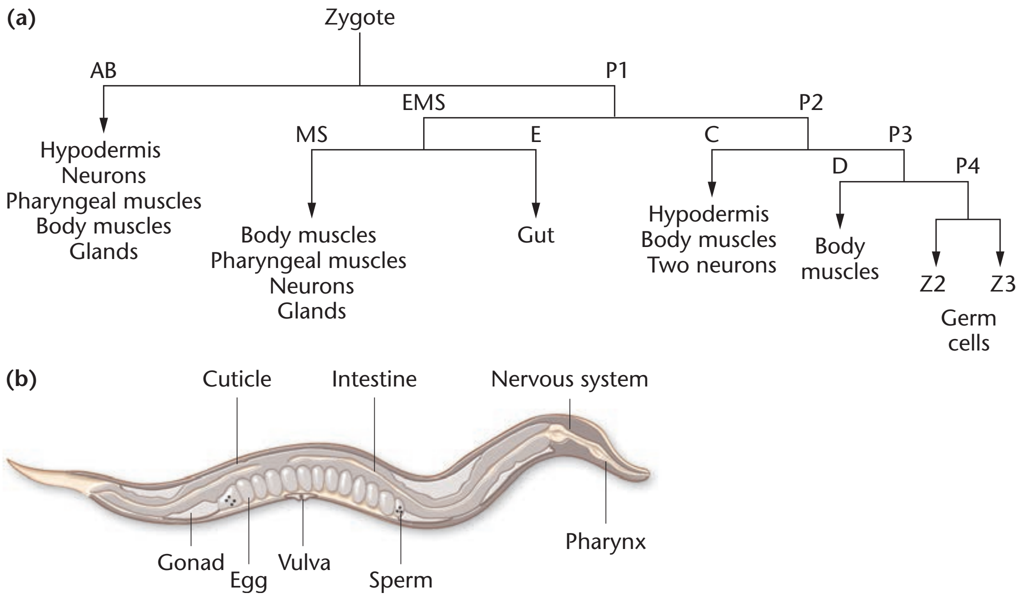


FIGURE 23.21 (a) A truncated cell lineage chart for *C. elegans*, showing early divisions and the tissues and organs formed from these lineages. Each vertical line represents a cell division, and horizontal lines connect the two cells produced. For example, the first division of the zygote creates two new cells, AB and P1. During embryogenesis, cell divisions will produce the 959 somatic cells of the adult hermaphrodite worm. (b) An adult *C. elegans* hermaphrodite. This nematode, about 1 mm in length, consists of 959 cells and is widely used as a model organism to study the genetic control of development.

life cycle includes an embryonic stage (about 16 hours), four larval stages (L1 through L4), and the adult stage. Adults are of two sexes: XX self-fertilizing hermaphrodites that can make both eggs and sperm, and XO males. Self-fertilization of mutagen-treated hermaphrodites is used to develop homozygous stocks of mutant strains, and hundreds of such mutants have been generated, cataloged, and mapped.

Adult hermaphrodites have 959 somatic cells (and about 2000 germ cells). The lineage of each cell, from fertilized egg to adult, has been mapped (Figure 23.21) and is invariant from individual to individual. Knowing the lineage of each cell, we can easily follow altered cell fates generated by mutations or by killing specific cells with laser microbeams or ultraviolet irradiation. In hermaphrodites, the developmental fate of cells in the reproductive system is determined by cell–cell interaction, illustrating how gene expression and cell–cell interaction work together to specify developmental outcomes.

Genetic Analysis of Vulva Formation

Adult *C. elegans* hermaphrodites lay eggs through the vulva, an opening near the middle of the body (Figure 23.21). The vulva is formed in stages during larval development and involves several rounds of cell–cell interactions.

In *C. elegans*, interaction between two neighboring cells, Z1.ppp and Z4.aaa, determines which will become the gonadal anchor cell (from which the vulva forms) and which will become a precursor to the uterus (**Figure 23.22**). The determination of which cell becomes which occurs during the second larval stage (L2) and is controlled by the Notch receptor gene, *lin-12*. In recessive *lin-12(0)* mutants (a loss-of-function mutant), no functional receptor protein is present, and both cells become anchor cells. The dominant mutation *lin-12(d)* (a gain-of-function mutation) causes both to become uterine precursors. Thus, expression of *lin-12* directs selection of the uterine pathway, because in the absence of the LIN-12 (Notch) receptor, both cells become anchor cells.

However, the situation is more complex than it first appears. Initially, the two neighboring cells are developmentally equivalent. Each synthesizes low levels of the Notch signal protein (encoded by the *lag-2* gene) *and* the Notch receptor protein. By chance, one cell ends up secreting more of the signal (LAG-2 or Delta protein) than the other cell. This causes the neighboring cell to increase production of the receptor (LIN-12 protein). The cell producing more of the receptor protein becomes the uterine precursor, and the other cell, producing more signal protein, becomes the anchor cell. The critical factor in this first round of cell–cell interaction is the balance between the LAG-2 (Delta) signal gene product and the LIN-12 (Notch) receptor gene product.

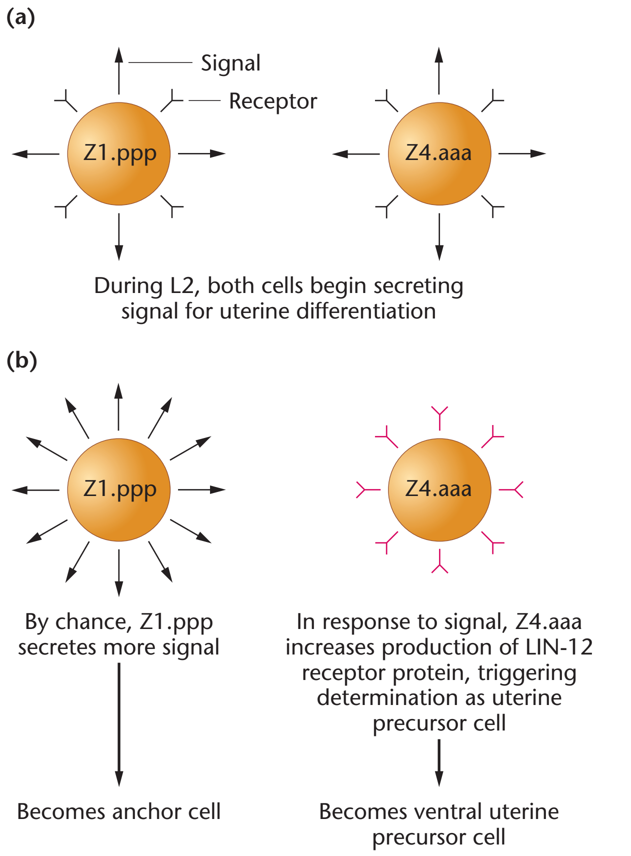

(a)
Signal
Receptor
Z1.ppp
Z4.aaa
During L2, both cells begin secreting signal for uterine differentiation
(b)
Z1.ppp
Z4.aaa
By chance, Z1.ppp secretes more signal
In response to signal, Z4.aaa increases production of LIN-12 receptor protein, triggering determination as uterine precursor cell
Becomes anchor cell
Becomes ventral uterine precursor cell

FIGURE 23.22 Cell–cell interaction in anchor cell determination. (a) During L2, two neighboring cells begin the secretion of chemical signals for the induction of uterine differentiation. (b) By chance, cell Z1.ppp produces more of these signals, causing cell Z4.aaa to increase production of the receptor for signals. The action of increased signals causes Z4.aaa to become the ventral uterine precursor cell and allows Z1.ppp to become the anchor cell.

Once the gonadal anchor cell has been determined, a second round of cell–cell interaction leads to formation of the vulva. This interaction involves the anchor cell (located in the gonad) and six neighboring cells (called precursor cells) located in the skin (**Figure 23.23**). The precursor cells, named P3.p to P8.p, are called Pn.p cells. The developmental fate of each Pn.p cell is specified by its position relative to the anchor cell.

During vulval development, the LIN-3 signal protein is synthesized by the anchor cell; this signal is received and processed by three adjacent Pn.p precursor cells (Pn.p 5–7). The cell closest to the anchor cell (usually Pn.p 6) becomes the primary vulval precursor cell, and the adjacent cells (Pn.p 5 and 7) become secondary precursor cells. A signal protein from the primary vulval precursor cell activates the *lin-12* receptor gene in the secondary cells, preventing them from becoming primary precursor cells. The other precursor cells (Pn.p 3, 4, and 8) receive no signal from the anchor cell and become skin cells.

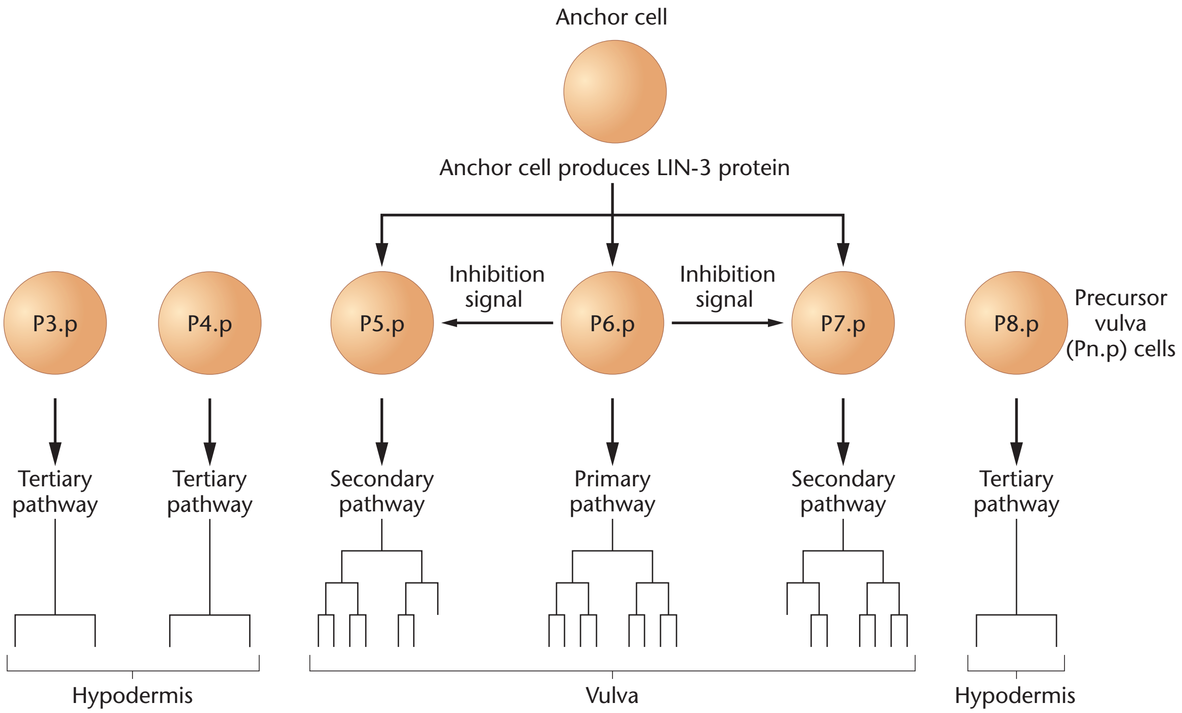


FIGURE 23.23 Cell lineage determination in *C. elegans* vulva formation. A signal from the anchor cell in the form of LIN-3 protein is received by three precursor cells (Pn.p cells). The cell closest to the anchor cell becomes the primary vulval precursor cell, and adjacent cells become secondary precursor cells. The primary cell produces a signal that activates the *lin-12* gene in secondary cells, preventing them from becoming primary cells. Flanking precursor cells, which receive no signal from the anchor cell, become skin (hypodermis) cells, instead of vulval cells.

23.8 Binary Switch Genes and Regulatory Networks Program Genomic Expression

During development, certain genes act as switches, decreasing the number of alternative developmental pathways that a cell can follow. Each decision point is usually binary—that is, there are two alternative developmental fates for a cell at a given time—and the action of a switch gene programs the cell to follow only one of these pathways. These genes are called **binary switch genes**. They are defined by their ability to initiate complete development of an organ or a tissue type, and in combination with signaling pathways, form **gene-regulatory networks (GRNs)**. We will briefly describe how a binary switch gene controls the formation of the eye and how this regulatory pathway is used in all organisms with eyes.

The Control of Eye Formation

Drosophila adults have compound eyes [Figure 23.24] that develop in preadult stages. Action of the wild-type alleles of the master binary switch genes *eyeless* and *twin of eyeless* program cells to follow the developmental pathway for eye formation instead of the pathway for antenna formation. In flies homozygous for the recessive mutant allele *eyeless*, the eyes are reduced in size with irregular facets or are absent altogether [Figure 23.24]. In developmental pathways that normally specify the formation of other organs such as legs, wings, and antennae, abnormal expression of *eyeless* results in eye formation on legs [Figure 23.25(a)] and other body parts. This indicates that expression of the *eyeless* gene during development of these structures overrides the normal program of determination and differentiation, causing cells to follow the developmental program for eye formation instead of the normal pathway.

The *eyeless* gene is part of a network of seven genes [Figure 23.25(b)], which is the master regulator of eye formation. As shown in the figure, this network is interconnected by feedback loops and is not a linear pathway. These genes encode transcription factors, which in turn regulate the expression of many other genes that control cell determination and specification during development of the eye. The key regulators in this network are *twin of eyeless* (*toy*) and *eyeless* (*ey*). The products of *toy* and *ey* initiate eye formation by activating the downstream transcription factors *sine oculis* (*so*) by binding to regulatory elements in an *so* enhancer. The So protein forms a dimer

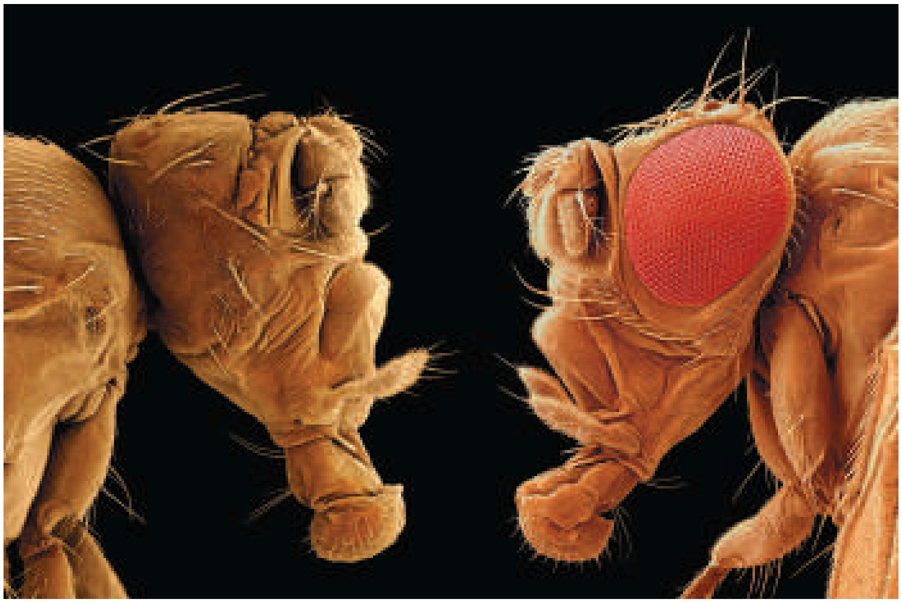


FIGURE 23.24 (Left) In flies homozygous for the *eyeless* (*ey*) mutation, eye development is abnormal and adults have no eyes. (Right) The normal compound eye of adult *Drosophila*. The *ey* gene is a binary switch gene regulating eye development in all animals.

with the protein encoded by the gene *eyes absent/clift* (*eya/cli*), which, in turn, activates the gene *dachshund* (*dac*) and starts the formation of feedback loops. The network also receives input from the genes *eyegone* (*eyg*) and *optix* (*opt*). During the stages of eye formation, this network interacts with several signaling pathways, including *Notch* (*N*), *hedgehog* (*hh*), and *decapentaplegic* (*dpp*), to form the retinal determination gene network (RDGN). Overall, the

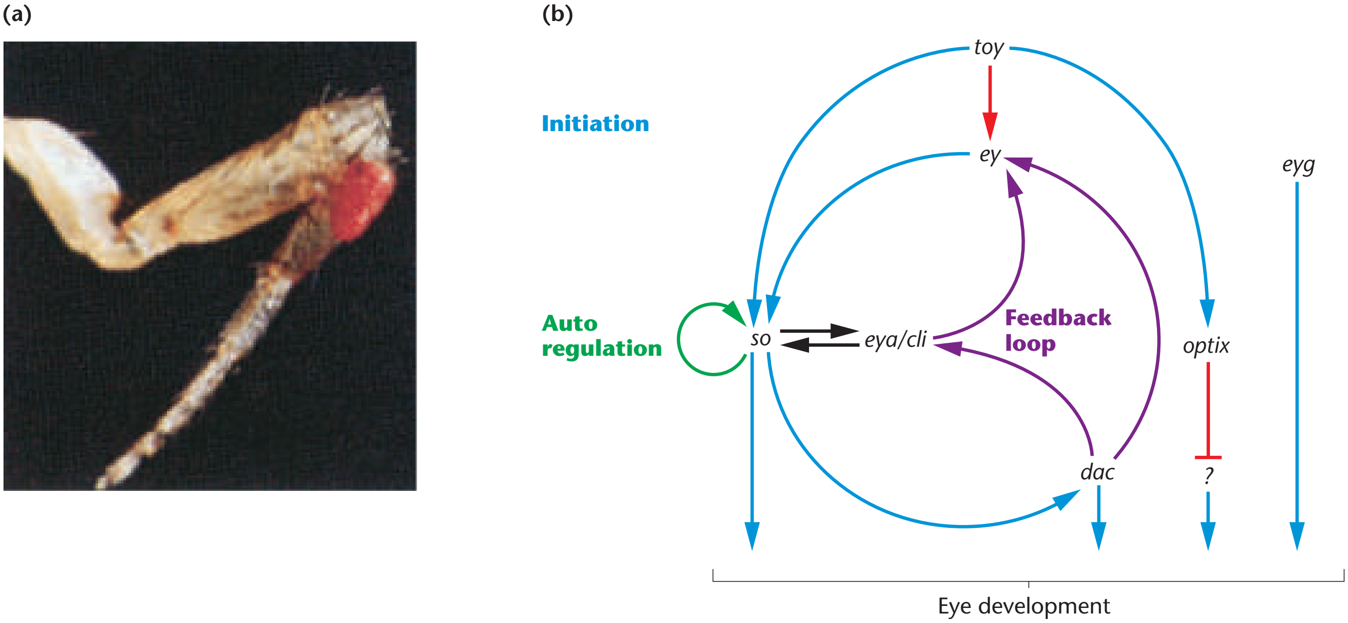


FIGURE 23.25 (a) Eye formation on the leg of a fly. This ectopic eye results from *eyeless* expression in cells normally destined to form a leg. (b) The genes *toy*, *ey* (and perhaps *eyg*) are master control genes and are at the top of the hierarchy of gene action in eye development. Other genes, *so*, *eya/cli* and *dac*, all of which encode transcription factors, are second-level genes that are regulated by the master control genes. This complex program is a network with genes interconnected via feedback loops, and not a linear system. Although activated independently, *eyg* expression is required for eye formation, and this gene acts cooperatively with *ey* in the developmental program.

TABLE 23.5 Eye Genes in *Drosophila* and Vertebrates

Vertebrate Gene	*Drosophila* Homolog	Expression in Vertebrate Eye	Loss of Function
Pax6	*eyeless* (*ey*), *twin of eyeless* (*toy*)	Lens placode, optic vesicle	Aniridia (human), *small eye* (mouse)
Bmp4	*Decapentaplegic* (*Dpp*)	Optic vesicle, head ectoderm	No lens (mouse)
Bmp7	*glass bottom boat* (*gbb*)	Optic vesicle, head ectoderm	No lens (mouse)
EYA4	*clift* (*cli*) [previously named *eyes absent* (*eya*)]	Perioptic mesenchyme, weak lens expression	No eye phenotype (human) or (mouse), some human mutations lead to cataracts and anterior defects
Six3	*sine oculis* (*so*)	Lens placode, optic vesicle	Very small eyes (human)
Optx2	*Optix* (*Optix*)	Optic vesicle	No eyes (human)
Dach1	*dachshund* (*dac*)	Optic vesicle	Not determined

Source: Wawersik, S., and Maas, R. L. (2000). Vertebrate eye development as modelled in *Drosophila*. *Hum. Mol. Genet.* 9:917–925, Table 1, p. 921.

activation of the seven genes in the eye developmental pathway may eventually involve expression of up to 1000 genes in the preadult and adult stages that participate in eye formation and vision.

The *eyeless* allele and the other genes in this network have been highly conserved during evolution and are used by all animals, including humans, to make eyes. The human and fly genes in this network are compared in **Table 23.5**. The discovery that *eyeless* directs the formation of eyes in vertebrates forced reevaluation of the long-held belief that the compound eye of insects and the single-lens eye of vertebrates evolved independently. This assumption was based on the observation that the compound insect eye and the vertebrate camera eye have different embryonic origins, develop by different pathways, and are structurally very different. However, at the molecular level, the *eyeless* gene and its vertebrate equivalent, *Pax6*, have a high degree of sequence homology in encoded regions critical for DNA binding, and both are expressed in the development of the eye in flies and humans.

Walter Gehring and his colleagues examined the relationship between *eyeless* and *Pax6* by generating transgenic *Drosophila* that carried copies of the mouse *Pax6* gene. The striking results led to two conclusions: (1) the vertebrate version of this binary switch gene was capable of triggering the formation of extra eyes on the antennae of flies [**Figure 23.26(a)**], and (2) because the mouse gene works in flies, the invertebrate and vertebrate eyes *are* in fact homologous at the molecular level. Therefore, the eyes of *Drosophila* and the mouse, and in fact, all animals with eyes, are related evolutionarily. The downstream targets of these transcription factors are also conserved, indicating that steps in the genetic control of eye development are shared between species that diverged over half a billion years ago from a common ancestor. This evolutionary conservation makes it possible to use genetic analysis in *Drosophila* to study the development of eyes and to explore the molecular basis for inherited eye defects in humans.

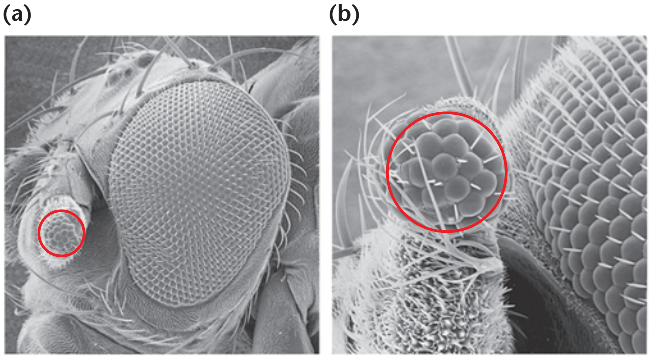


FIGURE 23.26 (a) Eye formation on the antenna (circled) of *Drosophila* induced by action of the mouse version of the *eyeless* gene (*Small eye*). (b) High magnification of the induced eye (circled) showing that the eye structure on the antenna is normal.

GENETICS, ETHICS, AND SOCIETY

Stem Cell Wars

The study of stem cells is one of the most promising and controversial areas of scientific research. All the cells that make up the approximately 200 distinct types of tissues in our bodies are descended from stem cells. Stem cells are undifferentiated and have the capacity to both replicate indefinitely and differentiate into cells with specialized functions, such as those of the heart, brain, liver, and muscle tissue. Some types of stem cells are defined as *totipotent*, meaning that they have the ability to differentiate into any mature cell type in the body, as well as tissues associated with the developing embryo, such as the placenta. Other types of stem cells are *pluripotent*, meaning that they are able to differentiate into any of a smaller number of mature cell types. In contrast, mature, fully *differentiated* cells do not replicate or undergo transformations into different cell types.

It is now possible to isolate and culture human pluripotent stem cells. These cells remain undifferentiated and grow indefinitely in culture dishes. When treated with growth factors or hormones, these cultured pluripotent stem cells can differentiate into cells of many mature types including neural, bone, kidney, liver, heart, or pancreatic cells.

The fact that pluripotent stem cells can differentiate into specialized cells has created great excitement and hope. Someday it may be possible to harvest unlimited numbers of specialized cells to replace those in damaged and diseased tissues. Hence, stem cells could be used to treat Parkinson disease, Type 1 diabetes, chronic heart disease, Alzheimer disease, and spinal cord injuries, as well as correct genetic defects and treat cancers. Given the potential of stem cell therapies, why should stem cell research be controversial?

The answer to that question lies in the source of the pluripotent stem cells. Until recently, all pluripotent stem cell lines were derived from five-day-old embryonic blastocysts. Blastocysts at this stage consist of 50–150 cells, most of which will develop into placental and extraembryonic tissues of the early embryo. The inner cell mass within the blastocyst consists of about 30–40 pluripotent stem cells that can develop into all the embryo's tissues. *In vitro* fertilization clinics routinely grow fertilized eggs to the five-day blastocyst stage prior to uterine transfer. Human embryonic stem cell (ESC) lines are created by taking the inner cell mass out of five-day blastocysts that are being discarded by fertilization clinics, and growing the cells in culture dishes.

The fact that early embryos are destroyed in the process of establishing human ESC lines disturbs people who believe that preimplantation embryos are persons with rights; however, it does not disturb people who believe that these embryos are too primitive to have the status of a human being. Both sides in the debate invoke fundamental questions of what constitutes a human being.

Recently, scientists have developed several other types of pluripotent stem cells that are not derived from embryos. One type—known as *induced pluripotent stem (iPS) cells*–uses adult somatic cells as the source of stem cell lines. These somatic cells are injected with engineered retroviruses that integrate into the cells' DNA. The retroviruses contain several cloned human genes that encode products responsible for converting the somatic cells into immortal, pluripotent stem cells. Another type, known as *stimulus-triggered acquisition of pluripotency (STAP) cells*, is derived from somatic cells that are treated with various types of stresses, such as low pH. These stresses convert the cells to pluripotency.

The development of iPS and STAP cell lines has generated increased enthusiasm for stem cell research, as these cells are thought to circumvent the ethical problems associated with the use of human embryos. In addition, they may become sources of patient-specific pluripotent stem cell lines that can be used for transplantation, without immune system rejection.

At the present time, it is unknown whether stem cell therapies of any type will be as miraculous as predicted; so the controversies and promises remain.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links dealing with the ethical and technological challenges surrounding stem cell research and therapies.

1. Despite the promise of pluripotent stem cell therapies and dozens of clinical trials worldwide, no treatments have yet been approved. What are some of the technical and ethical problems that create challenges for stem cell research?

   Read about these questions in Scudellari, M. (2016). A decade of iPS cells. *Nature* 534:310–312. *Also see* Kimmelman, J., et al. (2016). Global standards for stem-cell research. *Nature* 533:311–313.

2. Although many early clinical trials using ESC and iPS cells are under way, hundreds of unregulated clinics throughout the world offer stem cell treatments. Discuss the efforts to create guidelines and regulations governing these "rogue" clinics. What are the ethical arguments both for and against their existence?

   Unregulated stem cell therapies are discussed in A Closer Look at Stem Cells, *sponsored by the International Society for Stem Cell Research (ISSCR)* (http://www.closerlookatstemcells.org/).

CASE STUDY One foot or another

In humans, *HOXD* genes play a critical role in limb development. An analysis of several families with malformations known as rocker bottom foot [congenital vertical talus (CVT)] or claw foot [Charcot Marie Tooth (CMT)] revealed that CVT has an autosomal dominant form of inheritance with variable expressivity and incomplete penetrance. (See Chapter 4 for a discussion of penetrance and expressivity.) Genomic analysis identified a single missense mutation of the transcription factor *HOXD10* as the cause of both malformations. The mutation was located in the homeodomain recognition sequence that is critical in determining DNA binding to target genes expressed during limb development. Conditions with variable expressivity and incomplete penetrance pose a number of problems in helping family members make decisions about whether or not to have children. It is difficult to estimate the risk of an affected child being born to a parental "carrier" who has the mutant allele but does not express it. In addition, it is difficult to estimate both the type of malformation and the degree of expression that might occur in an affected child.

1. In preparation for meeting with an unaffected family member who is already pregnant, a genetic counselor turns to you as a developmental geneticist with two questions:

   a. How might a dominant mutation in a gene encoding a transcription factor cause a developmental malformation in some cases, but be nonpenetrant in others?

   b. How can variable expressivity result in two clinically different disorders from the same mutation?

2. The counselor plans to tell the family member that genetic testing is available to determine whether or not she carries the mutation. From an ethical point of view, what details about the possible outcomes should be offered if the family member is found to carry the mutation and wants to proceed with testing?

See Pauli, R., and Motulsky, A. (1981). Risk counseling in autosomal dominant disorders with undetermined penetrance. *J. Med. Genet.* 18:340–343.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Developmental genetics, which explores the mechanisms by which genetic information controls development and differentiation, is one of the major areas of study in biology. Geneticists are investigating this topic by isolating developmental mutations and identifying the genes involved in developmental processes.
2. During embryogenesis, the activity of specific genes is controlled by the internal environment of the cell, including localized cytoplasmic components. In flies, the regulation of early events is mediated by the maternal cytoplasm, which then influences zygotic gene expression. As development proceeds, both the cell's internal environment and its external environment become further altered by the presence of early gene products and communication with other cells.
3. In *Drosophila*, both genetic and molecular studies have confirmed that the egg contains information specifying the body plan of the larva and adult.
4. Extensive genetic analysis of embryonic development in *Drosophila* has led to the identification of maternal-effect genes whose products establish the anterior–posterior axis of the embryo. In addition, these maternal-effect genes activate sets of zygotic segmentation genes, initiating a cascade of gene regulation that ends with the determination of segment identity by the homeotic selector genes. These same gene sets control aspects of embryonic development in all bilateral animals, including humans.
5. Flower formation in *Arabidopsis* is controlled by homeotic genes, but these gene sets are from a different gene family than the homeotic selector genes of *Drosophila* and other animals.
6. In *C. elegans* and many other organisms, cell–cell signaling systems program the developmental fate of adjacent and distant cells.
7. During development, major transitions are often controlled by binary switch genes whose action results in the selection of one of two alternate pathways that results in the formation of organ or tissue types.

INSIGHTS AND SOLUTIONS

1. In the slime mold *Dictyostelium*, experimental evidence suggests that cyclic AMP (cAMP) plays a central role in the developmental program leading to spore formation. The genes encoding the cAMP cell-surface receptor have been cloned, and the amino acid sequence of the protein components is known. To form reproductive structures, free-living individual cells aggregate together and then differentiate into one of two cell types, prespore cells or prestalk cells. Aggregating cells secrete waves or oscillations of cAMP to foster the aggregation of cells and then continuously secrete cAMP to activate genes in the aggregated cells at later stages of development. It has been proposed that cAMP controls cell–cell interaction and gene expression. It is important to test this hypothesis by using several experimental techniques. What different approaches can you devise to test this hypothesis, and what specific experimental systems would you employ to test them?

   Solution: Two of the most powerful forms of analysis in biology involve the use of biochemical analogs (or inhibitors) to block gene transcription or the action of gene products in a predictable way, and the use of mutations to alter genes and their products. These two approaches can be used to study the role of cAMP in the developmental program of *Dictyostelium*. First, compounds chemically related to cAMP, such as GTP and GDP, can be used to test whether they have any effect on the processes controlled by cAMP. In fact, both

GTP and GDP lower the affinity of cell-surface receptors for cAMP, effectively blocking the action of cAMP.

Mutational analysis can be used to dissect components of the cAMP receptor system. One approach is to use transformation with wild-type genes to restore mutant function. Similarly, because the genes for the receptor proteins have been cloned, it is possible to construct mutants with known alterations in the component proteins and transform them into cells to assess their effects.

2. In the sea urchin, early development may occur even in the presence of actinomycin D, which inhibits RNA synthesis. However, if actinomycin D is present early in development but is removed a few hours later, all development stops. In fact, if actinomycin D is present only between the sixth and eleventh hours of development, events that normally occur at the fifteenth hour are arrested. What conclusions can be drawn concerning the role of gene transcription between hours 6 and 15?

Solution: Maternal mRNAs are present in the fertilized sea urchin egg. Thus, a considerable amount of development can take place without transcription of the embryo's genome. Because development past 15 hours is inhibited by prior treatment with actinomycin D, it appears that transcripts from the embryo's genome are required to initiate or maintain these events. This transcription must take place between the sixth and fifteenth hours of development.

3. If it were possible to introduce one of the homeotic genes from *Drosophila* into an *Arabidopsis* embryo homozygous for a homeotic flowering gene, would you expect any of the *Drosophila* genes to negate (rescue) the *Arabidopsis* mutant phenotype? Why or why not?

Solution: The *Drosophila* homeotic genes belong to the *Hox* gene family, whereas *Arabidopsis* homeotic genes belong to the MADS-box protein family. Both gene families are present in *Drosophila* and *Arabidopsis*, but they have evolved different functions in the animal and the plant kingdoms. As a result, it is unlikely that a transferred *Drosophila Hox* gene would rescue the phenotype of a MADS-box mutant, but only an actual experiment would confirm this.

Problems and Discussion Questions

1. **HOW DO WE KNOW?** In this chapter, we have focused on large-scale as well as the inter- and intracellular events that take place during embryogenesis and the formation of adult structures. In particular, we discussed how the adult body plan is laid down by a cascade of gene expression, and the role of cell—cell communication in development. Based on your knowledge of these topics, answer several fundamental questions:
(a) How have we discovered that specific genes control development in an organism like *Drosophila*?
(b) How do we know that molecular gradients in the egg of *Drosophila* exist?
(c) How did we discover that selector genes specify which adult structures will be formed by body segments?
(d) How did we learn about the levels of gene regulation involved in vulval development in *C. elegans*?
(e) How do we know that eye formation in all animals is controlled by a binary switch gene?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 555. These all center around concepts related to stages of development. Write a short essay based on these concepts that outlines the role of differential transcription, gene control of cell fate, and the role of signaling systems in development.

3. Carefully distinguish between the terms *differentiation* and *determination*. Which phenomenon occurs initially during development?

4. Nuclei from almost any source may be injected into *Xenopus* oocytes. Studies have shown that these nuclei remain active in transcription and translation. How can such an experimental system be useful in developmental genetic studies?

5. Distinguish between the syncytial blastoderm stage and the cellular blastoderm stage in *Drosophila* embryogenesis.

6. (a) What are maternal-effect genes?
(b) When are gene products from these genes made, and where are they located?
(c) What aspects of development do maternal-effect genes control?
(d) What is the phenotype of maternal-effect mutations?

7. (a) What are zygotic genes, and when are their gene products made?
(b) What is the phenotype associated with zygotic gene mutations?
(c) Does the maternal genotype contain zygotic genes?

8. List the main classes of zygotic genes. What is the function of each class of these genes?

9. Experiments have shown that any nuclei placed in the polar cytoplasm at the posterior pole of the *Drosophila* egg will differentiate into germ cells. If polar cytoplasm is transplanted into the anterior end of the egg just after fertilization, what will happen to nuclei that migrate into this cytoplasm at the anterior pole?

10. How can you determine whether a particular gene is being transcribed in different cell types?

11. You observe that a particular gene is being transcribed during development. How can you tell whether the expression of this gene is under transcriptional or translational control?

12. The homeotic mutation *Antennapedia* causes mutant *Drosophila* to have legs in place of antennae and is a dominant gain-of-function mutation. What are the properties of such mutations? How does the *Antennapedia* gene change antennae into legs?

13. The *Drosophila* homeotic mutation *spineless aristapedia* (ss^a) results in the formation of a miniature tarsal structure (normally part of the leg) on the end of the antenna. What insight is provided by (ss^a) concerning the role of genes during determination?

14. Embryogenesis and oncogenesis (generation of cancer) share a number of features including cell proliferation, apoptosis, cell migration and invasion, formation of new blood vessels, and differential gene activity. Embryonic cells are relatively undifferentiated, and cancer cells appear to be undifferentiated or

dedifferentiated. Homeotic gene expression directs early development, and mutant expression leads to loss of the differentiated state or an alternative cell identity. M. T. Lewis [(2000). *Breast Can. Res.* 2:158–169] suggested that breast cancer may be caused by the altered expression of homeotic genes. When he examined 11 such genes in cancers, 8 were underexpressed while 3 were overexpressed compared with controls. Given what you know about homeotic genes, could they be involved in oncogenesis?

15. Early development depends on the temporal and spatial interplay between maternally supplied material and mRNA and the onset of zygotic gene expression. Maternally encoded mRNAs must be produced, positioned, and degraded [Surdej and Jacobs-Lorena (1998). *Mol. Cell Biol.* 18:2892–2900]. For example, transcription of the *bicoid* gene that determines anterior–posterior polarity in *Drosophila* is maternal. The mRNA is synthesized in the ovary by nurse cells and then transported to the oocyte, where it localizes to the anterior ends of oocytes. After egg deposition, *bicoid* mRNA is translated and unstable bicoid protein forms a decreasing concentration gradient from the anterior end of the embryo. At the start of gastrulation, *bicoid* mRNA has been degraded. Consider two models to explain the degradation of *bicoid* mRNA: (1) degradation may result from signals within the mRNA (intrinsic model), or (2) degradation may result from the mRNA's position within the egg (extrinsic model). Experimentally, how could one distinguish between these two models?

16. Formation of germ cells in *Drosophila* and many other embryos is dependent on their position in the embryo and their exposure to localized cytoplasmic determinants. Nuclei exposed to cytoplasm in the posterior end of *Drosophila* eggs (the pole plasm) form cells that develop into germ cells under the direction of maternally derived components. R. Amikura et al. [(2001). *Proc. Nat. Acad. Sci. (USA)* 98:9133–9138] consistently found mitochondria-type ribosomes outside mitochondria in the germ plasma of *Drosophila* embryos and postulated that they are intimately related to germ-cell specification. If you were studying this phenomenon, what would you want to know about the activity of these ribosomes?

17. One of the most interesting aspects of early development is the remodeling of the cell cycle from rapid cell divisions, apparently lacking G1 and G2 phases, to slower cell cycles with measurable G1 and G2 phases and checkpoints. During this remodeling, maternal mRNAs that specify cyclins are deadenylated, and zygotic genes are activated to produce cyclins. Audic et al. [(2001). *Mol. and Cell. Biol.* 21:1662–1671] suggest that deadenylation requires transcription of zygotic genes. Present a diagram that captures the significant features of these findings.

Extra-Spicy Problems

18. A number of genes that control expression of *Hox* genes in *Drosophila* have been identified. One of these homozygous mutants is *extra sex combs*, where some of the head and all of the thorax and abdominal segments develop as the last abdominal segment. In other words, all affected segments develop as posterior segments. What does this phenotype tell you about which set of *Hox* genes is controlled by the *extra sex combs* gene?

19. The *apterous* gene in *Drosophila* encodes a protein required for wing patterning and growth. It is also known to function in nerve development, fertility, and viability. When human and mouse genes whose protein products closely resemble *apterous* were used to generate transgenic *Drosophila* [Rincon-Limas et al. (1999). *Proc. Nat. Acad. Sci. (USA)* 96:2165–2170], the *apterous* mutant phenotype was *rescued*. In addition, the whole-body expression patterns in the transgenic *Drosophila* were similar to normal *apterous*.

(a) What is meant by the term *rescued* in this context?

(b) What do these results indicate about the molecular nature of development?

20. In *Arabidopsis*, flower development is controlled by sets of homeotic genes. How many classes of these genes are there, and what structures are formed by their individual and combined expression?

21. The floral homeotic genes of *Arabidopsis* belong to the MADS-box gene family, while in *Drosophila*, homeotic genes belong to the homeobox gene family. In both *Arabidopsis* and *Drosophila*, members of the *Polycomb* gene family control expression of these divergent homeotic genes. How do *Polycomb* genes control expression of two very different sets of homeotic genes?

22. Vulval development in *C. elegans* is dependent on the response of some of the central epidermal progenitor cells in the region of the developing vulva to a chemical signal from the gonad. Signaling from the gonad is blocked by action of the vulvaless mutant *let-23* so that none of the central progenitor cells form vulval structures. In the vulvaless mutant, *n300*, the central progenitor cells do not form.

(a) Which gene is likely to act earlier in the vulval developmental pathway?

(b) What phenotype (vulva formed or vulvaless) would you expect from the double mutant? Why?

23. Much of what we know about gene interactions in development has been learned using nematodes, yeast, flies, and bacteria. This is due, in part, to the relative ease of genetic manipulation of these well-characterized genomes. However, of great interest are gene interactions involving complex diseases in humans. Wang and White [(2011). *Nature Methods* 8(4):341–346] describe work using RNAi to examine the interactive proteome in mammalian cells. They mention that knockdown inefficiencies and off-target effects of introduced RNAi species are areas that need particular improvement if the methodology is to be fruitful.

(a) How might one use RNAi to study developmental pathways?

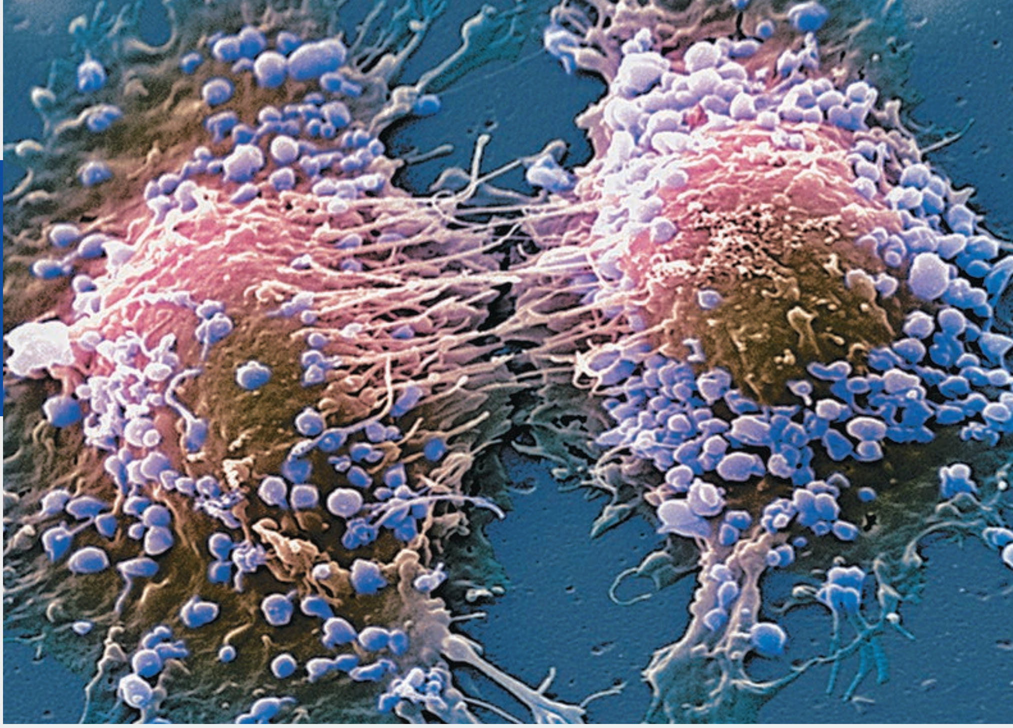
(b) Comment on how "knockdown inefficiencies" and "off-target effects" would influence the interpretation of results.

24. Dominguez et al. (2004) suggest that by studying genes that determine growth and tissue specification in the eye of *Drosophila*, much can be learned about human eye development.

(a) What evidence suggests that genetic eye determinants in *Drosophila* are also found in humans? Include a discussion of orthologous genes in your answer.

(b) What evidence indicates that the *eyeless* gene is part of a developmental network?

24



Cancer Genetics

Colored scanning electron micrograph of two prostate cancer cells in the final stages of cell division (cytokinesis). The cells are still joined by strands of cytoplasm.

CHAPTER CONCEPTS

- Cancer is characterized by genetic defects in fundamental aspects of cellular function, including DNA repair, chromatin modification, cell-cycle regulation, apoptosis, and signal transduction.
- Most cancer-causing mutations occur in somatic cells; only about 5 to 10 percent of cancers have a hereditary component.
- Cancer cells in primary and secondary tumors are clonal, arising from one stem cell that accumulated several cancer-causing mutations.
- The development of cancer is a multistep process requiring multiple mutations and clonal expansions.
- Cancer cells show high levels of genomic instability, leading to the accumulation of multiple mutations, some in cancer-related genes.
- Epigenetic effects such as DNA methylation and histone modifications play significant roles in the development of cancers.
- Mutations in proto-oncogenes and tumor-suppressor genes contribute to the development of cancers.
- Genetic changes that lead to cancer can be triggered by virus infections as well as numerous natural and human-made carcinogens.

Cancer is the second most common cause of death in Western countries, exceeded only by heart disease. It strikes people of all ages, and one out of three people will experience a cancer diagnosis sometime in his or her lifetime. Each year, more than 14 million cases of cancer are diagnosed worldwide and more than 8 million people will die from the disease.

Over the last 40 years, scientists have discovered that cancer is a genetic disease at the somatic cell level, characterized by the presence of gene products derived from mutated or abnormally expressed genes. The combined effects of numerous abnormal gene products lead to the uncontrolled growth and spread of cancer cells. Although some mutated cancer genes may be inherited, most are created within somatic cells that then divide and form tumors. Completion of the Human Genome Project and numerous large-scale rapid DNA-sequencing studies have opened the door to a wealth of new information about the mutations that trigger a cell to become cancerous. This new understanding of cancer genetics is also leading to new gene-specific treatments, some of which are now entering clinical trials. Some scientists predict that gene-targeted therapies will replace chemotherapies within the next 25 years.

The goal of this chapter is to highlight our current understanding of the nature and causes of cancer. As we will see, cancer is a genetic disease that arises from the accumulation of mutations in genes controlling many basic aspects of cellular function. We will examine the relationship between genes and cancer and consider how mutations, chromosomal changes, epigenetics, and environmental agents play roles in the development of cancer. Please note that some of the topics discussed in this chapter are explored in greater depth elsewhere in the text (see Chapter 19 and Special Topic 3—Genomics and Precision Medicine).

24.1 Cancer Is a Genetic Disease at the Level of Somatic Cells

Perhaps the most significant development in understanding the causes of cancer is the realization that cancer is a genetic disease. Genomic alterations that are associated with cancer range from single-nucleotide substitutions to large-scale chromosome rearrangements, amplifications, and deletions (Figure 24.1). However, unlike other genetic diseases, cancer is caused by mutations that arise predominantly in somatic cells. Only about 5 to 10 percent of cancers are associated with germ-line mutations that increase a person's susceptibility to certain types of cancer. Another important difference between cancers and other genetic diseases is that cancers rarely arise from a single mutation in a single gene. They arise instead from the accumulation of many mutations in many genes. The mutations that lead to cancer affect multiple cellular functions, including repair of DNA damage, cell division, apoptosis, cellular differentiation, migratory behavior, and cell—cell contact.

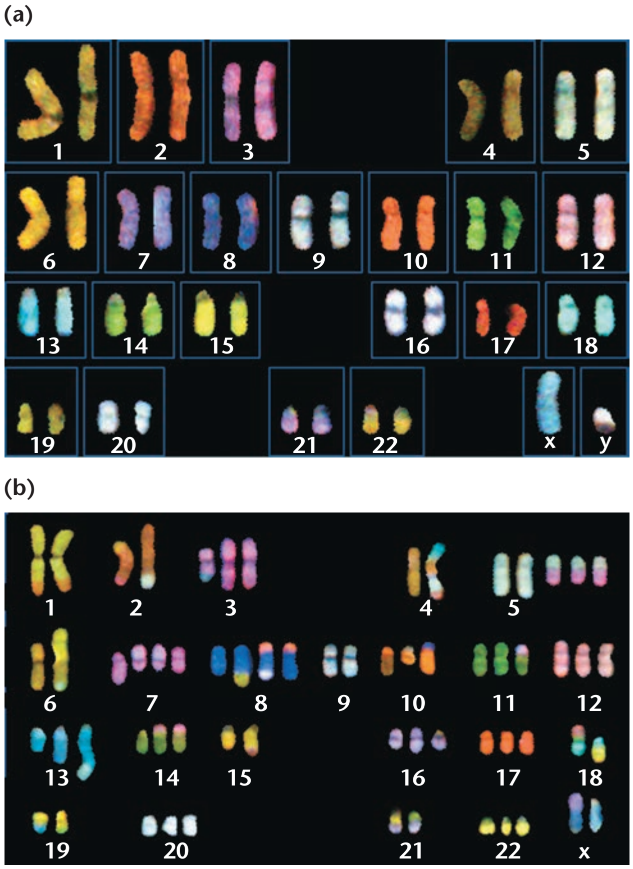


FIGURE 24.1 (a) Spectral karyotype of a normal cell. (b) Karyotype of a cancer cell showing translocations, deletions, and aneuploidy—characteristic features of cancer cells.

What Is Cancer?

Clinically, cancer defines a large number of complex diseases that behave differently depending on the cell types from which they originate and the types of genetic alterations that occur within each cancer type. Cancers vary in their ages of onset, growth rates, invasiveness, prognoses, and responsiveness to treatments. However, at the molecular level, all cancers exhibit common characteristics that unite them as a family.

All cancer cells share two fundamental properties: (1) abnormal cell growth and division (**proliferation**), and (2) defects in the normal restraints that keep cells from spreading and colonizing other parts of the body (**metastasis**). In normal cells, these functions are tightly controlled by genes that are expressed appropriately in time and place. In cancer cells, these genes are either mutated or are expressed inappropriately.

It is this combination of uncontrolled cell proliferation and metastatic spread that makes cancer cells dangerous. When a cell simply loses genetic control over cell growth, it may grow into a multicellular mass, a **benign tumor**. Such a tumor can often be removed by surgery and may cause no serious harm. However, if cells in the tumor also have the ability to break loose, enter the bloodstream, invade other tissues, and form secondary tumors (**metastases**), they become malignant. **Malignant tumors** are often difficult to treat and may become life threatening. As we will see later in the chapter, there are multiple steps and genetic mutations that convert a benign tumor into a dangerous malignant tumor.

The Clonal Origin of Cancer Cells

Although malignant tumors may contain billions of cells, and may invade and grow in numerous parts of the body, all cancer cells in the primary and secondary tumors are clonal, meaning that they originated from a common ancestral cell that accumulated specific cancer-causing mutations. This is an important concept in understanding the molecular causes of cancer and has implications for its diagnosis.

Numerous data support the concept of cancer clonality. For example, reciprocal chromosomal translocations are characteristic of many cancers, including leukemias and lymphomas (two cancers involving white blood cells). Cancer cells from patients with Burkitt lymphoma show reciprocal translocations between chromosome 8 (with translocation breakpoints at or near the *c-myc* gene) and chromosomes 2, 14, or 22 (with translocation breakpoints at or near one of the immunoglobulin genes). Each patient with Burkitt lymphoma exhibits unique breakpoints in his or her *c-myc* and immunoglobulin gene DNA sequences; however, all lymphoma cells within that patient contain identical translocation breakpoints. This demonstrates that cells in a tumor arise from a single cell, and this cell passes on its genetic aberrations to its progeny.

Although all cancer cells within a tumor are clonal, containing the same core set of cancer-causing genes that arose in the ancestral tumor cell, not all cells in a tumor are genetically identical throughout their entire genomes. Next-generation sequencing studies reveal that tumors are composed of subpopulations, or subclones, each of which contains its own sets of distinctive mutations. We will discuss the origins and implications of cancer subclones later in the chapter.

Driver Mutations and Passenger Mutations

Scientists are now applying some of the recent advances in DNA sequencing to identify all the somatic mutations within tumors. These studies compare the DNA sequences of genomes from cancer cells and normal cells derived from the same patient. Data from these studies are revealing that tens of thousands of somatic mutations can be present in cancer cells. Researchers believe that only a handful of mutations in each tumor—called **driver mutations**—give a growth advantage to a tumor cell. The remainder of the mutations may be acquired over time, perhaps as a result of the high levels of DNA damage that occurs in cancer cells, but these mutations have no direct contribution to the cancer phenotype. These are known as **passenger mutations**. The total number of driver mutations that occur in any particular cancer is small—between 2 and 8.

It is now possible to sequence the genomes of individual tumor cells. These studies confirm that there is a great deal of genetic variation between individual cells and subclones within tumors. Most of these variations are due to the accumulation of different types of passenger mutations, with the few key driver mutations remaining constant between subclones. Although most of these passenger mutations may not initially confer a selective advantage on the cells that contain them, if environmental conditions change—such as during chemotherapy or radiotherapy—a passenger mutation may confer a new phenotype such as drug resistance which will be selected for, leading to clonal expansion of that cell and the appearance of a new subclone within the tumor.

As we will discover in subsequent sections of this chapter, the genes that acquire driver mutations that lead to cancer (called oncogenes and tumor-suppressor genes) are those that control a large number of essential cellular functions including DNA damage repair, chromatin modification, cell-cycle regulation, and programmed cell death.

The Cancer Stem Cell Hypothesis

A concept that is related to the clonal origin of cancer cells is that of the cancer stem cell. Many scientists now believe that most of the cells within tumors do not proliferate. Those that do proliferate and give rise to all the cells within the tumor or within a tumor subclone are known as **cancer stem cells**. Stem cells are undifferentiated cells that have the capacity for self-renewal—a process in which the stem cell divides unevenly, creating one daughter cell that goes on to differentiate into a mature cell type and one that remains a stem cell. Stem cells are also discussed elsewhere in the text (see the Genetics, Ethics, and Society essay in Chapter 23). The cancer stem cell hypothesis is in contrast to the random or stochastic model that predicts that every cell within a tumor has the potential to form a new tumor.

Cancer stem cells have been identified in leukemias as well as in solid tumors of the brain, breast, colon, ovary, pancreas, and prostate. It is still not clear what fraction of any tumor is composed of cancer stem cells. For example, human acute myeloid leukemias contain less than 1 cancer stem cell in 10,000. In contrast, some solid tumors may contain as many as 40 percent cancer stem cells.

It is possible that cancer stem cells may arise from normal adult stem cells within a tissue, or they may be created from more differentiated somatic cells that acquire properties similar to stem cells after accumulating numerous mutations and changes to chromatin structure.

Cancer as a Multistep Process, Requiring Multiple Mutations and Clonal Expansions

Although we know that cancer is a genetic disease initiated by driver mutations that lead to uncontrolled cell proliferation and metastasis, a single mutation is not sufficient to transform a normal cell into a tumor-forming (tumorigenic) malignant cell. If it were sufficient, then cancer would be far more prevalent than it is. In humans, mutations occur spontaneously at a rate of about 10^{-6} mutations per gene, per cell division, mainly due to the intrinsic error rates of DNA replication. Because there are approximately 10^{16} cell divisions in a human body during a lifetime, a person might suffer up to 10^{10} mutations per gene somewhere in the body, during his or her lifetime. However, only about one person in three will suffer from cancer.

The phenomenon of age-related cancer is another indication that cancer develops from the accumulation of several mutagenic events in a single cell. The incidence of most cancers rises exponentially with age. If a single mutation were sufficient to convert a normal cell to a malignant one, then cancer incidence would appear to be independent of age. Another indication that cancer is a multistep process is the delay that occurs between exposure to **carcinogens** (cancer-causing agents) and the appearance of the cancer. For example, there was an incubation period of five to eight years between the time people were exposed to radiation from the atomic explosions at Hiroshima and Nagasaki and the onset of leukemias.

Each step in **tumorigenesis** (the development of a malignant tumor) appears to be the result of two or more genetic alterations that release a cancer stem cell from the controls that normally operate on proliferation and malignancy. Each step confers a selective advantage to the growth and survival of the cell and is propagated through successive **clonal expansions** leading to a fully malignant tumor.

The stepwise clonal evolution of tumors is illustrated by the development of colorectal cancer. Colorectal cancers are known to proceed through several clinical stages that are characterized by the stepwise accumulation of genetic defects in several genes (**Figure 24.2**). The first step is the conversion of a normal epithelial cell into a small cluster of cells known as an adenoma or polyp. This step requires inactivating mutations in the adenomatous polyposis coli (APC) gene, a gene that encodes a protein involved in the normal differentiation of intestinal cells. The *APC* gene is a tumor-suppressor gene, which will be discussed later in the chapter. The resulting adenoma grows slowly and is considered benign.

The second step in the development of colorectal cancer is the acquisition of a second genetic alteration in one of the cells within the small adenoma. This is usually a mutation in the *Kras* gene, a gene whose product is normally involved with regulating cell growth. The mutations in *Kras* that contribute to colorectal cancer cause the Kras protein to become constitutively active, resulting in unregulated cell division. The cell containing the *APC* and *Kras* mutations grows by clonal expansion to form a larger intermediate adenoma of approximately 1 cm in diameter. The cells of the original small adenoma (containing the *APC* mutation) are now vastly outnumbered by cells containing the two mutations.

The third step, which transforms a large adenoma into a malignant tumor (carcinoma), requires several more waves of clonal expansions triggered by the acquisition of defects in several genes, including *TP53*, *PI3K*, and *TGF*-β. The products of these genes control several important aspects of normal cell growth and division, such as apoptosis, growth signaling, and cell-cycle regulation—all of which we will discuss in more detail later in the chapter. The resulting carcinoma is able to further grow and invade the underlying tissues of the colon. A few cells within the carcinoma may break free of the tumor, migrate to other parts of the body, and form metastases.

24.2 Cancer Cells Contain Genetic Defects Affecting Genomic Stability, DNA Repair, and Chromatin Modifications

Cancer cells contain large numbers of mutations and chromosomal abnormalities. Many researchers believe that the fundamental defect in cancer cells is a derangement of the cells' normal ability to repair DNA damage. The resulting loss of genomic integrity leads to a general increase in the mutation rate for every gene in the genome, including cancer-causing driver mutations. The high level of genomic instability seen in cancer cells is known as the **mutator phenotype**. In addition, recent research has revealed that cancer cells contain aberrations in the types and locations of chromatin modifications, particularly DNA and histone methylation patterns.

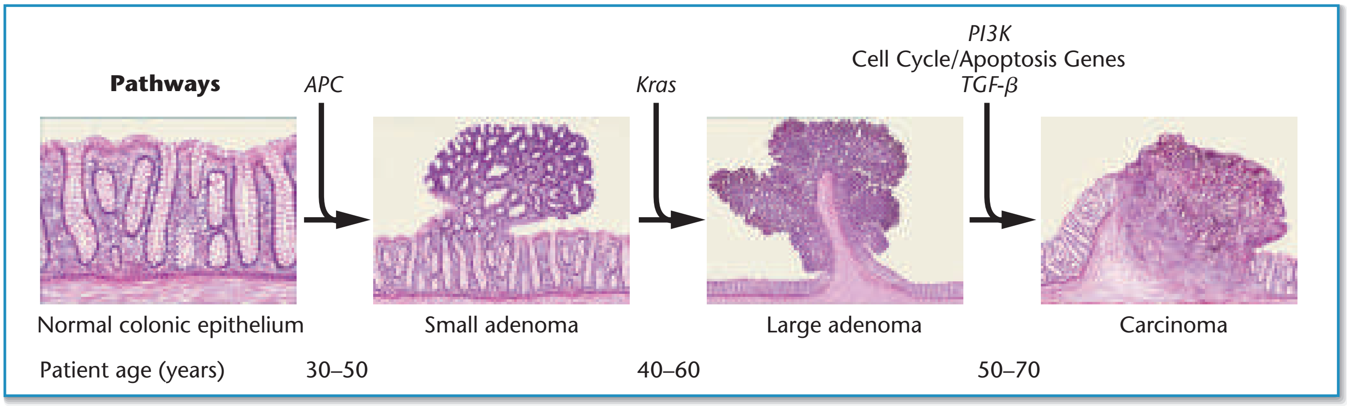


FIGURE 24.2 Steps in the development of colorectal cancers. Some of the genes that acquire driver mutations and cause the progressive development of colorectal cancer are shown above the photographs. These driver mutations accumulate over time and can take 40 years or more to result in the formation of a malignant tumor.

Genomic Instability and Defective DNA Repair

Genomic instability in cancer cells is characterized by the presence of somatic point mutations and chromosomal effects such as translocations, aneuploidy, chromosome loss, DNA amplification, and deletions (Figure 24.1 and Figure 24.3). Cancer cells that are grown in cultures in the lab also show a great deal of genomic instability—duplicating, losing, and translocating chromosomes or parts of chromosomes. Often cancer cells show specific chromosomal defects that are used to diagnose the type and stage of the cancer. For example, leukemic white blood cells from patients with chronic myelogenous leukemia (CML) bear a specific translocation, in which the *C-ABL* gene on chromosome 9 is translocated into the *BCR* gene on chromosome 22. This translocation creates a structure known as the **Philadelphia chromosome** (Figure 24.4). The *BCR-ABL* fusion gene codes for a chimeric BCR-ABL protein. The normal ABL protein is a protein kinase that acts within signal transduction pathways, transferring growth factor signals from the external environment to the nucleus. The BCR-ABL protein is an abnormal signal transduction molecule in CML cells, which stimulates these cells to proliferate even in the absence of external growth signals.

In keeping with the concept of the cancer mutator phenotype, a number of inherited cancers are caused by defects in genes that control DNA repair. For example, xeroderma pigmentosum (XP) is a rare hereditary disorder that is characterized by extreme sensitivity to ultraviolet (UV) light and other carcinogens. Patients with XP often develop skin cancer. Cells from patients with XP are defective in nucleotide excision repair, with mutations appearing in any one of seven genes whose products are necessary to carry out DNA repair. XP cells are impaired in their ability to repair DNA lesions such as thymine dimers induced by UV light. The relationship between XP and genes controlling nucleotide excision repair is also described earlier in the text (see Chapter 15).

Another example is hereditary nonpolyposis colorectal cancer (HNPCC), which is caused by mutations in genes controlling DNA repair. HNPCC is an autosomal dominant syndrome, affecting about 1 in every 200 to 1000 people. Patients affected by HNPCC have an increased risk of developing colon, ovary, uterine, and kidney cancers. Cells from patients with HNPCC show higher than normal mutation rates and genomic instability. At least eight genes are associated with HNPCC, and four of these genes control aspects of DNA mismatch repair. Inactivation of any of these four genes—*MSH2, MSH6, MLH1,* and *MLH3*—causes a rapid accumulation of genome-wide mutations and the subsequent development of cancers.

The observation that hereditary defects in genes controlling nucleotide excision repair and DNA mismatch repair lead to high rates of cancer lends support to the idea that the mutator phenotype is a significant contributor to the development of cancer.

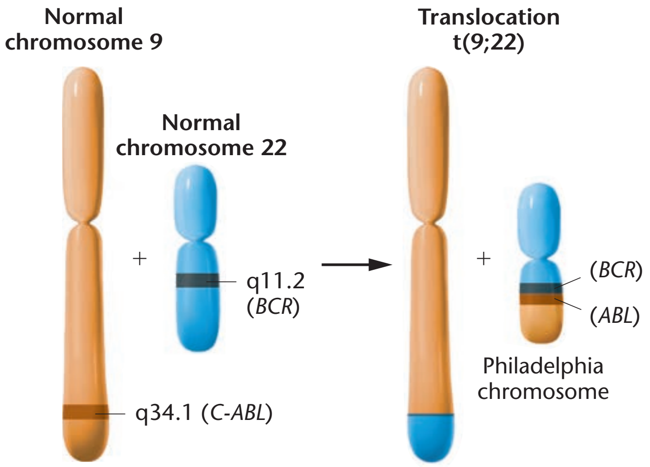


FIGURE 24.4 A reciprocal translocation involving the long arms of chromosomes 9 and 22 results in the formation of a characteristic chromosome, the Philadelphia chromosome, which is associated with chronic myelogenous leukemia (CML). The t(9;22) translocation results in the fusion of the *C-ABL* proto-oncogene on chromosome 9 with the *BCR* gene on chromosome 22. The fusion protein is a powerful hybrid molecule that allows cells to escape control of the cell cycle, contributing to the development of CML.

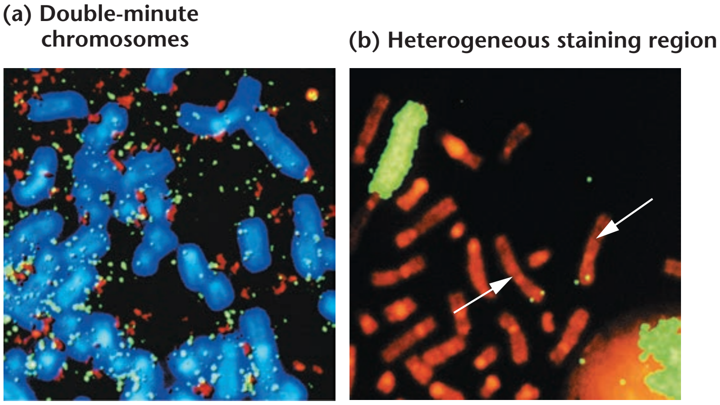

(a) Double-minute chromosomes
(b) Heterogeneous staining region

FIGURE 24.3 DNA amplifications in neuroblastoma cells. (a) Two cancer genes (*MYCN* in red and *MDM2* in green) are amplified as small DNA fragments that remain separate from chromosomal DNA within the nucleus. These units of amplified DNA are known as double-minute chromosomes. Normal chromosomes are stained blue. (b) Multiple copies of the *MYCN* gene are amplified within one large region called a heterogeneous staining region (green). Single copies of the *MYCN* gene are visible as green dots at the ends of the normal parental chromosomes (white arrows). Normal chromosomes are stained red.

Chromatin Modifications and Cancer Epigenetics

The field of cancer epigenetics is providing new perspectives on the genetics of cancer. **Epigenetics** is the study of chromosome-associated changes that affect gene expression but do not alter the nucleotide sequence of DNA. Epigenetic effects can be inherited from one cell to its progeny cells and may be present in either somatic or germ-line cells. DNA methylation and histone modifications such as acetylation and phosphorylation are examples of epigenetic modifications. The genomic patterns and locations of these modifications can affect gene expression. For example, DNA methylation is thought to be responsible for the gene silencing associated with parental imprinting, heterochromatin gene repression, and X chromosome inactivation. The effects of chromatin modifications and epigenetic factors on gene expression and cancer are discussed in more detail earlier in the text (see Chapters 17 and 19).

Cancer cells contain altered DNA methylation patterns. Overall, there is much less DNA methylation in cancer cells than in normal cells. At the same time, the promoters of some genes are hypermethylated in cancer cells. These changes are thought to result in the release of transcription repression over the bulk of genes that would be silent in normal cells—including cancer-causing genes—while at the same time repressing transcription of genes that would regulate normal cellular functions such as DNA repair and cell-cycle control.

Histone modifications are also disrupted in cancer cells. Genes that encode histone acetylases, deacetylases, methyltransferases, and demethylases are often mutated or aberrantly expressed in cancer cells. The large numbers of epigenetic abnormalities in tumors have prompted some scientists to speculate that there may be more epigenetic defects in cancer cells than there are gene mutations. In addition, because epigenetic modifications are reversible, it may be possible to treat cancers using epigenetic-based therapies.

Now Solve This

24.1 In chronic myelogenous leukemia (CML), leukemic blood cells can be distinguished from other cells of the body by the presence of a functional BCR-ABL hybrid protein. Explain how this characteristic provides an opportunity to develop a therapeutic approach to a treatment for CML.

- **Hint:** *This problem asks you to imagine a therapy that is based on the unique genetic characteristics of CML leukemic cells. The key to its solution is to remember that the BCR-ABL fusion protein is found only in CML white blood cells and that this unusual protein has a specific function thought to directly contribute to the development of CML. To help you answer this problem, you may wish to learn more about the cancer drug Gleevec (see https://www.cancer.gov/about-cancer/treatment/drugs/imatinibmesylate).*

24.3 Cancer Cells Contain Genetic Defects Affecting Cell-Cycle Regulation

One of the fundamental aberrations in all cancer cells is a loss of control over cell proliferation. Cell proliferation is the process of cell growth and division that is essential for all development and tissue repair in multicellular organisms. Although some cells, such as epidermal cells of the skin or blood-forming cells in the bone marrow, continue to grow and divide throughout an organism's lifetime, most cells in adult multicellular organisms remain in a nondividing, quiescent, and differentiated state. **Differentiated cells** are those that are specialized for specific functions, such as photoreceptor cells of the retina or muscle cells of the heart. The most extreme examples of nonproliferating cells are nerve cells, which divide little, if at all, even to replace damaged tissue. In contrast, many differentiated cells, such as those in the liver and kidney, are able to grow and divide when stimulated by extracellular signals and growth factors. In this way, multicellular organisms are able to replace dead and damaged tissue. However, the growth and differentiation of cells must be strictly regulated; otherwise, the integrity of organs and tissues would be compromised by the presence of inappropriate types and quantities of cells. Normal regulation over cell proliferation involves a large number of gene products that control steps in the cell cycle.

In this section, we will review steps in the cell cycle, some of the genes that control the cell cycle, and how these genes, when mutated, lead to cancer.

The Cell Cycle and Signal Transduction

The cellular events that occur in sequence from one cell division to the next comprise the **cell cycle** (**Figure 24.5**). The **interphase** stage of the cell cycle is the interval between mitotic divisions. During this time, the cell grows and replicates its DNA. During **G1**, the cell prepares for DNA synthesis by accumulating the enzymes and molecules required for DNA replication. G1 is followed by **S phase**, during which the cell's chromosomal DNA is replicated. During **G2**, the cell continues to grow and prepare for division. During **M phase**, the duplicated chromosomes condense, sister chromosomes separate to opposite poles, and the cell divides in two. These phases of the cell cycle are also discussed in more detail earlier in the text (see Chapter 2).

In early to mid-G1, the cell makes a decision either to enter the next cell cycle or to withdraw from the cell cycle into quiescence. Continuously dividing cells do not exit the cell cycle but proceed through G1, S, G2, and M phases; however, if the cell receives signals to stop growing, it enters the **G0** phase of the cell cycle. During G0, the cell remains

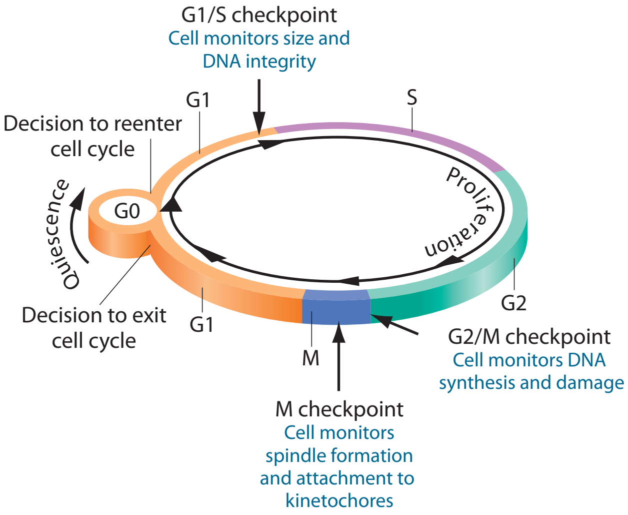

G1/S checkpoint
Cell monitors size and DNA integrity
G1
S
Decision to reenter cell cycle
Quiescence
G0
Proliferation
G2
Decision to exit cell cycle
G1
M
G2/M checkpoint
Cell monitors DNA synthesis and damage
M checkpoint
Cell monitors spindle formation and attachment to kinetochores

FIGURE 24.5 Checkpoints and proliferation decision points monitor the progress of the cell through the cell cycle.

metabolically active but does not grow or divide. Most differentiated cells in multicellular organisms can remain in this G0 phase indefinitely. Some, such as neurons, never reenter the cell cycle. In contrast, cancer cells are unable to enter G0, and instead, they continuously cycle. Their *rate* of proliferation is not necessarily any greater than that of normal proliferating cells; however, they are not able to become quiescent at the appropriate time or place.

Cells in G0 can often be stimulated to reenter the cell cycle by external growth signals. These signals are delivered to the cell by molecules such as growth factors and hormones that bind to cell-surface receptors, which then relay the signal from the plasma membrane to the cytoplasm. The process of transmitting growth signals from the external environment to the cell nucleus is known as **signal transduction**. Ultimately, signal transduction initiates a program of gene expression that propels the cell out of G0 back into the cell cycle. Cancer cells often have defects in signal transduction pathways. Sometimes, abnormal signal transduction molecules send continuous growth signals to the nucleus even in the absence of external growth signals. An example of abnormal signal transduction due to mutations in the *ras* gene is described in Section 24.4. In addition, malignant cells may not respond to external signals from surrounding cells—signals that would normally inhibit cell proliferation within a mature tissue.

Cell-Cycle Control and Checkpoints

In normal cells, progress through the cell cycle is tightly regulated, and each step must be completed before the next step can begin. There are at least three distinct points in the cell cycle at which the cell monitors external signals and internal equilibrium before proceeding to the next stage. These are the *G1/S*, *G2/M*, and *M checkpoints* (Figure 24.5). At the G1/S checkpoint, the cell monitors its size and determines whether its DNA has been damaged. If the cell has not achieved an adequate size, or if the DNA has been damaged, further progress through the cell cycle is halted until these conditions are corrected. If cell size and DNA integrity are normal, the G1/S checkpoint is traversed, and the cell proceeds to S phase. The second important checkpoint is the G2/M checkpoint, where physiological conditions in the cell are monitored prior to mitosis. If DNA replication or repair of any DNA damage has not been completed, the cell cycle arrests until these processes are complete. The third major checkpoint occurs during mitosis and is called the M checkpoint. At this checkpoint, both the successful formation of the spindle-fiber system and the attachment of spindle fibers to the kinetochores associated with the centromeres are monitored. If spindle fibers are not properly formed or attachment is inadequate, mitosis is arrested.

In addition to regulating the cell cycle at checkpoints, the cell controls progress through the cell cycle by means of two classes of proteins: *cyclins* and *cyclin-dependent kinases (CDKs)*. The cell accumulates and destroys cyclin proteins in a precise pattern during the cell cycle (**Figure 24.6**). When a cyclin is present, it binds to a specific CDK, triggering activity of the CDK/cyclin complex. The CDK/cyclin complex then selectively phosphorylates and activates other proteins that in turn bring about the changes necessary to advance the

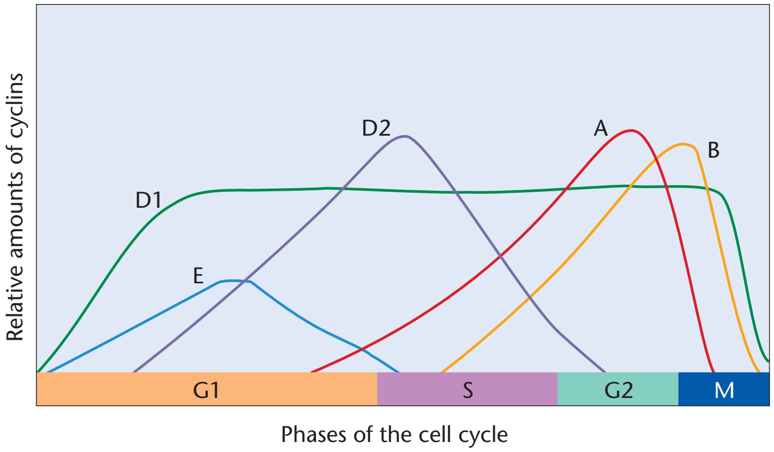


FIGURE 24.6 Relative expression times and amounts of cyclins during the cell cycle. Cyclin D1 accumulates early in G1 and is expressed at a constant level through most of the cycle. Cyclin E accumulates in G1, reaches a peak, and declines by mid–S phase. Cyclin D2 begins accumulating in the last half of G1, reaches a peak just after the beginning of S, and then declines by early G2. Cyclin A appears in late G1, accumulates through S phase, peaks at the G2/M transition, and is rapidly degraded. Cyclin B peaks at the G2/M transition and declines rapidly in M phase.

cell through the cell cycle. For example, in G1 phase, CDK4/cyclin D complexes activate proteins that stimulate transcription of genes whose products (such as DNA polymerase δ and DNA ligase) are required for DNA replication during S phase. Another CDK/cyclin complex, CDK1/cyclin B, phosphorylates a number of proteins that bring about the events of early mitosis, such as nuclear membrane breakdown, chromosome condensation, and cytoskeletal reorganization (Figure 24.7). Mitosis can only be completed, however, when cyclin B is degraded and the protein phosphorylations characteristic of M phase are reversed. Although a large number of different protein kinases exist in cells, only a few are involved in cell-cycle regulation.

Mutation or misexpression of any of the genes controlling the cell cycle can contribute to the development of cancer. For example, if genes that control the G1/S or G2/M checkpoints are mutated, the cell may continue to grow and divide without repairing DNA damage. As these cells continue to divide, they accumulate mutations in genes whose products control cell proliferation or metastasis. Similarly, if genes that control progress through the cell cycle, such as those that encode the cyclins, are expressed at the wrong time or at incorrect levels, the cell may grow and divide continuously and may be unable to exit the cell cycle into G0. The result in both cases is that the cell loses control over proliferation and is on its way to becoming cancerous.

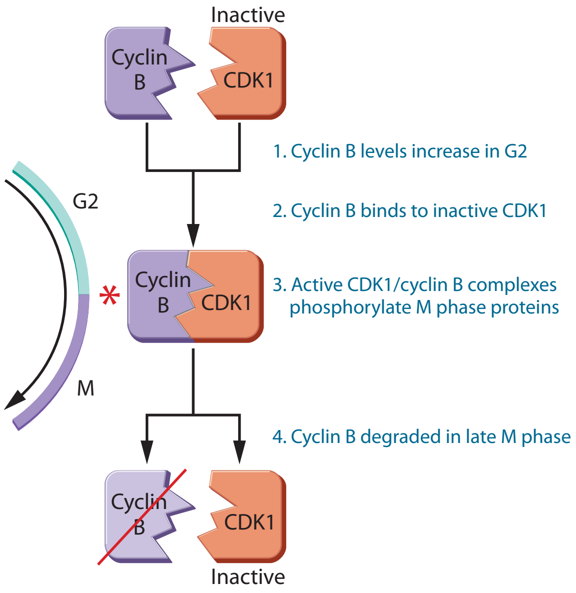

Inactive
Cyclin B
CDK1
1. Cyclin B levels increase in G2
G2
2. Cyclin B binds to inactive CDK1
*
Cyclin B
CDK1
3. Active CDK1/cyclin B complexes phosphorylate M phase proteins
M
4. Cyclin B degraded in late M phase
Cyclin B
CDK1
Inactive

FIGURE 24.7 CDK1 and cyclin B control the transition from G2 to M phase. In late G2 phase, cyclin B accumulates and forms complexes with inactive CDK1 molecules. CDK1 is activated within the complexes and adds phosphate groups to cellular components. These phosphorylated molecules bring about the structural and biochemical changes that are necessary for M phase. In late M phase, cyclin B is degraded, CDK1 becomes inactive, and the M phase phosphorylations are reversed.

Control of Apoptosis

As already described, if DNA replication, repair, or chromosome assembly is defective, normal cells halt their progress through the cell cycle until the condition is corrected. This reduces the number of mutations and chromosomal abnormalities that accumulate in normal proliferating cells. However, if DNA or chromosomal damage is so severe that repair is impossible, the cell may initiate a second line of defense—a process called **apoptosis**, or *programmed cell death*. Apoptosis is a genetically controlled process whereby the cell commits suicide. Besides its role in preventing cancer, apoptosis is also initiated during normal multicellular development to eliminate certain cells that do not contribute to the adult organism. The steps in apoptosis are the same for damaged cells and for cells being eliminated during development: nuclear DNA becomes fragmented, internal cellular structures are disrupted, and the cell dissolves into small spherical structures known as apoptotic bodies Figure 24.8. In the final step, the immune system's phagocytic cells engulf the apoptotic bodies. A series of proteases called **caspases** are responsible for initiating apoptosis and for digesting intracellular components.

Apoptosis is genetically controlled in that regulation of the levels and activities of specific gene products such as Bcl2 and BAX can trigger or prevent apoptosis. Members of the BAX group of proteins initiate apoptosis, and members of the Bcl2 group inhibit apoptosis. In normal cells, Bcl2 proteins inhibit the activity of BAX proteins, allowing the cell to survive. When the cell is under stresses such as DNA damage, Bcl2 activity is inhibited, allowing BAX to trigger apoptosis. Cancer cells often contain high levels of Bcl2 or reduced levels of BAX, preventing cells from entering apoptosis.

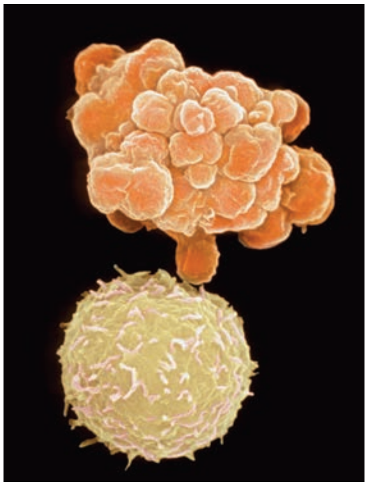


FIGURE 24.8 Normal white blood cell (bottom) and a white blood cell undergoing apoptosis (top). Apoptotic bodies appear as grape-like clusters on the cell surface.

Some of the same genes that control cell-cycle checkpoints can initiate apoptotic pathways. These genes are mutated in many cancers. As a result of the mutation or inactivation of these checkpoint genes, the cell is unable to either repair its DNA or undergo apoptosis. This leads to the accumulation of even more mutations in genes that control growth, division, and metastasis.

Cancer Therapies and Cancer Cell Biology

Two major treatments for cancer are *chemotherapy* and *radiation therapy* (*radiotherapy*). These treatments work by preferentially targeting cells that are proliferating. The agents used in these treatments damage many cellular components, including DNA, leading to apoptosis and other forms of cell death such as necrosis. Cancer cells are more susceptible to these agents than are normal cells because they are proliferating rapidly and are less efficient at repairing damaged DNA.

Chemotherapeutic agents attack many different aspects of cancer cell biology. For example, antihormone drugs such as those for tumors of the breast or prostate affect signal transduction pathways, thereby blocking cell growth. Some chemotherapeutic agents target mitotic cells, preventing chromosomes from segregating. For example, anti-microtubule drugs such as vinblastine prevent microtubule formation, resulting in cell-cycle arrest during mitosis and destruction of arrested cells by apoptosis. Alkylating agents covalently bind proteins, RNA and DNA, creating interstrand and intrastrand crosslinks in DNA, which inhibit both DNA replication and transcription. Antimetabolites interfere with DNA or RNA synthesis by incorporating into replicating molecules or by inhibiting DNA synthesis enzymes. Topoisomerase inhibitors prevent double-stranded DNA unwinding during transcription and DNA replication. In all cases, chemotherapies cause cell damage, which in turn brings about cell death by various mechanisms including apoptosis.

Radiation therapies also work by preferentially damaging proliferating cells. The most commonly used radiotherapies are X rays, gamma rays, and particle radiations such as neutron beams. Radiation damages DNA by causing single-strand and double-strand DNA breaks. If damage is extensive, it may lead to apoptosis or other forms of cell death.

Radiotherapies and chemotherapies can often cause side effects. Most are due to the damage these agents inflict on normal cells that are also proliferating, such as blood-forming cells or cells that line the intestines. Normal proliferating cells are more sensitive to these agents than are normal nonproliferating cells; however, normal cells usually have intact DNA repair systems which can minimize their susceptibility to these therapies.

24.4 Proto-oncogenes and Tumor-Suppressor Genes Are Altered in Cancer Cells

Two general categories of genes are mutated or misexpressed in cancer cells—the proto-oncogenes and the tumor-suppressor genes (**Table 24.1**). **Proto-oncogenes** encode transcription factors that stimulate expression of other genes, signal transduction molecules that stimulate cell division, and cell-cycle regulators that move the cell through the cell cycle. Their products are important for normal cell functions, especially cell growth and division. When normal cells become quiescent and cease division, they repress the expression of most proto-oncogenes or modify the activities of their products. In cancer cells, one or more

TABLE 24.1 Some Proto-oncogenes and Tumor-Suppressor Genes

	Normal Function	Alteration in Cancer	Associated Cancers
Proto-oncogene			
c-myc	Transcription factor, regulates cell cycle, differentiation, apoptosis	Translocation, amplification, point mutations	Lymphomas, leukemias, lung cancer, many types
c-kit	Tyrosine kinase, signal transduction	Mutation	Sarcomas
RARα	Hormone-dependent transcription factor, differentiation	Chromosomal translocations with *PML* gene, fusion product	Acute promyelocytic leukemia
E6	Human papillomavirus encoded oncogene, inactivates p53	HPV infection	Cervical cancer
Cyclins	Bind to CDKs, regulate cell cycle	Gene amplification, overexpression	Lung, esophagus, many types
Tumor-Suppressor			
RB1	Cell-cycle checkpoints, binds E2F	Mutation, deletion, inactivation by viral oncogene products	Retinoblastoma, osteosarcoma, many types
TP53	Transcription regulation	Mutation, deletion, viruses	Many types
BRCA1, BRCA2	DNA repair	Point mutations	Breast, ovarian, prostate cancers

proto-oncogenes are altered in such a way that the activities of their products cannot be regulated in a normal fashion. This is sometimes due to mutations that result in an abnormal protein product. In other cases, proto-oncogenes may be overexpressed or expressed at an incorrect time due to mutations within gene-regulatory regions such as enhancer elements or due to alterations in chromatin structure that affect gene expression. If a proto-oncogene is continually in an "on" state, its product may constantly stimulate the cell to divide. When a proto-oncogene is mutated or abnormally expressed and contributes to the development of cancer, it is known as an **oncogene**—a cancer-causing gene. Oncogenes are proto-oncogenes that have experienced a gain-of-function alteration. As a result, only one allele of a proto-oncogene needs to be mutated or misexpressed to contribute to cancer. Hence, oncogenes confer a dominant cancer phenotype.

Tumor-suppressor genes are genes whose products normally regulate cell-cycle checkpoints or initiate the process of apoptosis. In normal cells, proteins encoded by tumor-suppressor genes halt progress through the cell cycle in response to DNA damage or growth-suppression signals from the extracellular environment. When tumor-suppressor genes are mutated or inactivated, cells are unable to respond normally to cell-cycle checkpoints or are unable to undergo programmed cell death if DNA damage is extensive. This leads to the accumulation of more mutations and the development of cancer. When both alleles of a tumor-suppressor gene are inactivated through mutation or epigenetic modifications, and other changes in the cell keep it growing and dividing, cells may become tumorigenic.

The following are examples of proto-oncogenes and tumor-suppressor genes that contribute to cancer when mutated or abnormally expressed. Genome-wide sequencing studies of cancer cells have identified approximately 200 oncogenes and tumor-suppressor genes, and more will likely be discovered as cancer research continues.

The *ras* Proto-oncogenes

Some of the most frequently mutated genes in human tumors are those in the ***ras* gene family**. These genes are mutated in more than 30 percent of human tumors. The *ras* gene family encodes signal transduction molecules that are associated with the cell membrane and regulate cell growth and division. Ras proteins normally transmit signals from the cell membrane to the nucleus, stimulating the cell to divide in response to external growth factors (Figure 24.9). Ras proteins alternate between an inactive (switched off) and an active (switched on) state by binding either guanosine diphosphate (GDP) or guanosine triphosphate (GTP). When a cell encounters a growth factor (such as platelet-derived growth factor or epidermal growth factor), growth factor receptors on the cell membrane bind to the growth factor, resulting in autophosphorylation of the cytoplasmic portion of the growth factor receptor. This causes recruitment of proteins known as nucleotide exchange factors to the plasma membrane. These nucleotide exchange factors cause Ras to release GDP and bind GTP, thereby activating Ras. The active, GTP-bound form of Ras then sends its signals through cascades of protein phosphorylations in the cytoplasm. The end-point of these cascades is activation of nuclear transcription factors that stimulate expression of genes whose products drive the cell from quiescence into the cell cycle. Once Ras has sent its signals to the nucleus, it hydrolyzes GTP to GDP and becomes inactive. Mutations that convert the *ras* proto-oncogene to an oncogene prevent the Ras protein from hydrolyzing GTP to GDP and hence freeze the Ras protein into its "on" conformation, constantly stimulating the cell to divide.

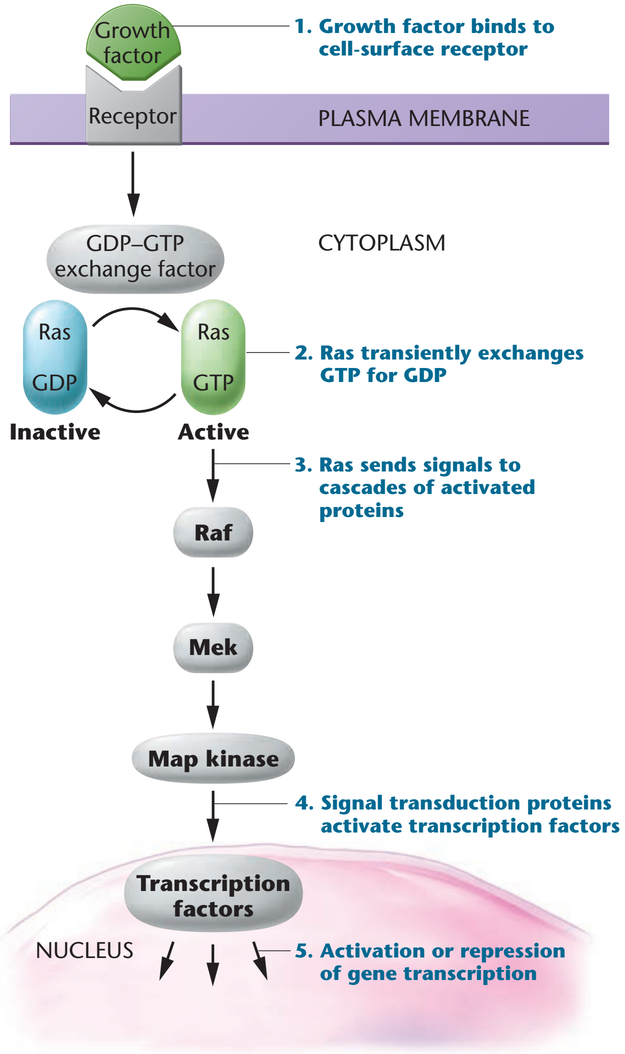

Growth factor
1. Growth factor binds to cell-surface receptor
Receptor
PLASMA MEMBRANE
GDP–GTP exchange factor
CYTOPLASM
Ras
GDP
Ras
GTP
2. Ras transiently exchanges GTP for GDP
Inactive
Active
3. Ras sends signals to cascades of activated proteins
Raf
Mek
Map kinase
4. Signal transduction proteins activate transcription factors
Transcription factors
NUCLEUS
5. Activation or repression of gene transcription

FIGURE 24.9 A signal transduction pathway mediated by Ras.

The *TP53* Tumor-Suppressor Gene

The most frequently mutated gene in human cancers—mutated in more than 50 percent of all cancers —is the *TP53* gene. This gene encodes a transcription factor (p53) that represses or stimulates transcription of more than 50 different genes.

Normally, the p53 protein is continuously synthesized but is rapidly degraded and therefore is present in cells at low levels. In addition, the p53 protein is normally bound to another protein called MDM2, which has several effects on p53. The presence of MDM2 on the p53 protein tags p53 for degradation and sequesters the transcriptional activation domain of p53. It also prevents the phosphorylations and acetylations that convert the p53 protein from an inactive to an active form.

Several types of cellular stress events bring about rapid increases in the nuclear levels of activated p53 protein. These include chemical damage to DNA, double-stranded breaks in DNA induced by ionizing radiation, and the presence of DNA-repair intermediates generated by exposure of cells to ultraviolet light. In response to these signals, MDM2 dissociates from p53, making p53 more stable and unmasking its transcription activation domain. Increases in the levels of activated p53 protein also result from increases in protein phosphorylation, acetylation, and other posttranslational modifications (**Figure 24.10**). Activated p53 protein acts as a transcription factor that stimulates expression of the *MDM2* gene. As the levels of MDM2 increase, p53 protein is again bound by MDM2, returned to an inactive state, and targeted for degradation, in a negative feedback loop.

The activated p53 protein initiates several different responses to DNA damage including cell-cycle arrest followed by DNA repair and apoptosis if DNA cannot be repaired. These responses are accomplished by p53 acting as a transcription factor that stimulates or represses the expression of genes involved in each response.

In normal cells, activated p53 can arrest the cell cycle at the G1/S and G2/M checkpoints, as well as retard the progression of the cell through S phase. To arrest the cell cycle at the G1/S checkpoint, activated p53 protein stimulates transcription of a gene encoding the p21 protein. The p21 protein inhibits the CDK4/cyclin D1 complex, hence preventing the cell from moving from G1 phase into S phase. Activated p53 protein also regulates expression of genes that retard the progress of DNA replication, thus allowing time for DNA damage to be repaired during S phase. By regulating expression of other genes, activated p53 can block cells at the G2/M checkpoint, if DNA damage occurs during S phase.

Activated p53 can also instruct a damaged cell to commit suicide by apoptosis. It does so by activating the transcription of the *BAX* gene and repressing transcription of the *BCL2* gene. As described previously, BAX proteins are important positive regulators of apoptosis and Bcl2 proteins are negative regulators. In cancer cells that lack functional p53, BAX protein levels do not increase in response to cell damage, Bcl2 levels remain high, and apoptosis may not occur. This increases the number of cells that survive with damaged DNA, leading to more mutations in proto-oncogenes and tumor-suppressor genes.

Although the majority of *TP53* mutations inactivate the p53 protein, several mutations confer a gain of function.

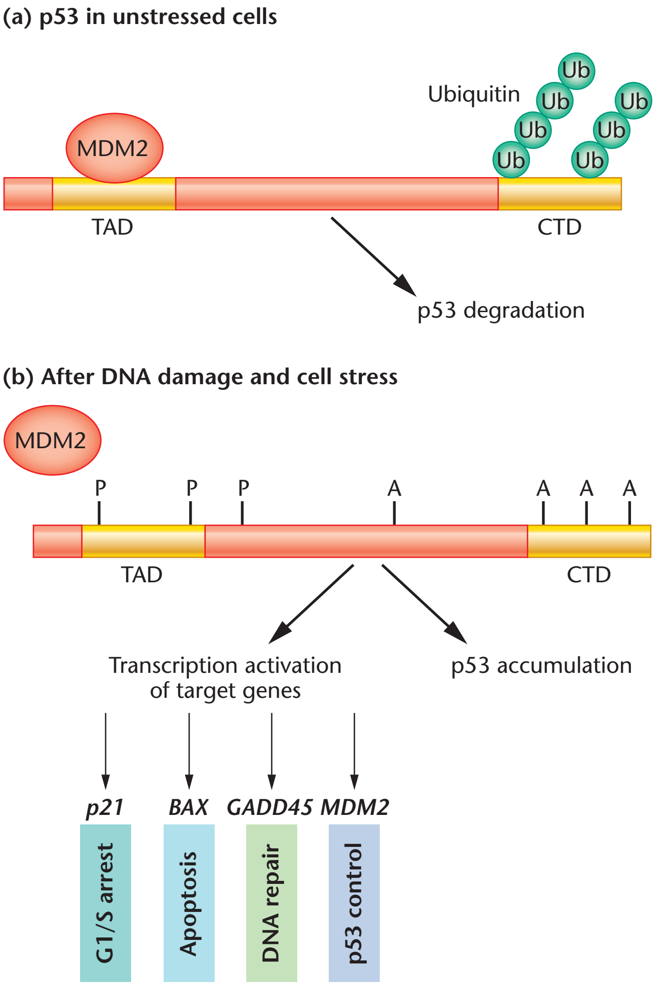


FIGURE 24.10 Steps in the regulation of p53 levels and activity. (a) In normal unstressed cells, p53 is kept inactive and at low abundance by MDM2, which binds to the transactivation domain (TAD) and stimulates the addition of ubiquitin onto lysine residues in the carboxy-terminal domain (CTD). The presence of ubiquitin promotes p53 degradation. (b) After various types of cellular stress, cellular kinases add phosphates (P's) to serines and threonines in the TAD, leading to dissociation of MDM2 and subsequent loss of ubiquitin. As the levels of p53 increase in the nucleus, acetyl transferases add acetyl groups (A's) to lysines in the CTD, which increases p53 stability and affinity for specific DNA sequences within the promoter regions of target genes. Examples of genes that are transcriptionally stimulated by p53 are *p21* (leading to G1/S cell-cycle arrest), *BAX* (stimulating apoptosis), *GADD45* (contributing to DNA repair), and *MDM2* (returning p53 to an inactive and low abundance state).

In these cases, mutant p53 increases the transcription of several genes whose products affect chromatin modifications, leading to genome-wide changes in histone methylation and acetylation and altered gene expression.

In summary, cells lacking functional p53 are unable to arrest at cell-cycle checkpoints or to enter apoptosis in response to DNA damage. As a result, they move unchecked through the cell cycle, regardless of the condition of the cell's DNA. This leads to high mutation rates and accumulation of mutations that lead to cancer. In addition, some mutated p53 proteins alter genome-wide patterns of chromatin modifications. Because of the importance of the *TP53* gene to genomic integrity, it is often referred to as the "guardian of the genome."

Now Solve This

24.2 People with a genetic condition known as Li–Fraumeni syndrome inherit one mutant copy of the *TP53* gene. These people have a high risk of developing a number of different cancers, such as breast cancer, leukemia, bone cancer, adrenocortical tumors, and brain tumors. Explain how mutations in one cancer-related gene can give rise to such a diverse range of tumors.

- **Hint:** *This problem involves an understanding of how tumor-suppressor genes regulate cell growth and behavior. The key to its solution is to consider which cellular functions are regulated by the p53 protein and how the absence of p53 could affect each of these functions. Also, read about loss of heterozygosity in Section 24.6.*

For more practice, see Problems 9, 10, and 25.

24.5 Cancer Cells Metastasize and Invade Other Tissues

As discussed at the beginning of this chapter, uncontrolled growth alone is insufficient to create a life-threatening cancer. Cancer cells must also become malignant, acquiring the ability to disengage from the original tumor site, to enter the blood or lymphatic system, to invade surrounding tissues, and to develop into secondary tumors. To leave the site of the primary tumor and invade other tissues, tumor cells secrete proteases that digest components of the extracellular matrix and basal lamina which are composed of proteins and carbohydrates. They surround and separate body tissues, form the scaffold for tissue growth, and inhibit the migration of cells. The ability to invade the extracellular matrix is also a property of some normal cell types. For example, implantation of the embryo in the uterine wall during pregnancy requires cell migration across the extracellular matrix. In addition, white blood cells reach sites of infection by penetrating capillary walls. The mechanisms of invasion are probably similar in these normal cells and in cancer cells.

Metastatic tumors arise from one or several cancer stem cells within one or more subclones of the primary tumor. Once these cells have disengaged from the primary tumor and traversed tissue barriers, they enter the blood or lymphatic system. Only a small percentage of circulating cancer cells—about 0.01 percent—survive to establish metastatic tumors. Once a metastasis is established, its cells continue to mutate and undergo clonal selections and expansions, similarly to those that occurred in the primary tumor.

Metastasis is controlled by a large number of gene products, including cell-adhesion molecules, cytoskeleton regulators, and proteolytic enzymes. For example, epithelial tumors have a lower than normal level of the E-cadherin glycoprotein, which is responsible for cell—cell adhesion in normal tissues. Also, proteolytic enzymes such as metalloproteinases are present at higher than normal levels in many highly malignant tumors. For example, breast cancer cells that metastasize to bone abnormally express the metalloproteinase gene *MMP1*. Those that spread to the lungs overexpress the *MMP1* and *MMP2* genes. It has been shown that the level of aggressiveness of a tumor correlates positively with the levels of proteolytic enzymes expressed by the tumor. In addition, malignant cells are not susceptible to the normal controls conferred by regulatory molecules such as tissue inhibitors of metalloproteinases (TIMPs).

24.6 Predisposition to Some Cancers Can Be Inherited

Although the vast majority of human cancers are sporadic, a small fraction (approximately 5 to 10 percent) have a hereditary or familial component. At present, more than 50 forms of hereditary cancer are known (**Table 24.2**).

Most inherited cancer-susceptibility alleles occur in tumor-suppressor genes and, though transmitted in a Mendelian dominant fashion, are not sufficient in themselves to trigger development of a cancer. Usually, at least one other somatic mutation in the other copy of the gene must occur to contribute to tumorigenesis. In addition, mutations in other genes are usually necessary to fully express the cancer phenotype. For example, inherited mutations in the *RB1* tumor-suppressor gene predispose individuals to developing various cancers including retinoblastoma. Although the normal somatic cells of these patients are heterozygous for the *RB1* mutation, cells within their tumors contain mutations in both copies of the gene or loss of the wild-type gene

TABLE 24.2 Some Inherited Predispositions to Cancer

Tumor Predisposition Syndrome	Chromosome	Gene Affected
Early-onset familial breast cancer	17q	*BRCA1*
Familial adenomatous polyposis	5q	*APC*
Familial melanoma	9p	*CDKN2*
Gorlin syndrome	9q	*PTCH1*
Hereditary nonpolyposis colon cancer	2p	*MSH2, 6*
Li-Fraumeni syndrome	17p	*TP53*
Multiple endocrine neoplasia, type 1	11q	*MEN1*
Multiple endocrine neoplasia, type 2	10q	*RET*
Neurofibromatosis, type 1	17q	*NF1*
Neurofibromatosis, type 2	22q	*NF2*
Retinoblastoma	13q	*pRb*
Von Hippel—Lindau syndrome	3p	*VHL*
Wilms tumor	11p	*WT1*

allele. The phenomenon whereby the second, wild-type, allele is lost is known as **loss of heterozygosity**. This can occur through chromosome deletions or rearrangements. Although mutation or loss of heterozygosity is an essential first step in expression of these inherited cancers, further mutations in other proto-oncogenes, tumor-suppressor genes, or chromatin-modifying genes are necessary for the tumor cells to become fully malignant.

In the study of hereditary cancers, those genes that bear germ-line mutations that are associated with an increased risk of hereditary cancers are sometimes called cancer predisposition genes. In most cases, these genes overlap with known or suspected proto-oncogenes and tumor-suppressor genes that suffer somatic mutations in noninherited cancers.

The development of hereditary colon cancer illustrates how inherited mutations in only one allele of a tumor-suppressor gene can contribute to malignancy. In Section 24.1, we described how colorectal cancers develop through the accumulation of mutations in several genes, leading to a stepwise clonal expansion of cells and the development of carcinomas. Although the vast majority of colorectal cancers are sporadic, about 1 percent of cases result from a genetic predisposition to cancer known as familial adenomatous polyposis (FAP). In FAP, individuals inherit one mutant copy of the *APC* (adenomatous polyposis) gene located on the long arm of chromosome 5. Mutations include deletions, frameshift, and point mutations. The normal function of the *APC* gene product is to act as a tumor suppressor controlling growth and differentiation. The presence of a heterozygous *APC* mutation causes the epithelial cells of the colon to partially escape cell-cycle control, and the cells divide to form small clusters of cells called polyps or adenomas. People who are heterozygous for this condition develop hundreds to thousands of colon and rectal polyps early in life. Although it is not necessary for the second allele of the *APC* gene to be mutated in polyps at this stage, in the majority of cases, the second *APC* allele becomes mutant or lost in a later stage of cancer development. The remaining steps in development of colorectal carcinoma follow the same order as that shown in Figure 24.2.

Now Solve This

24.3 Although tobacco smoking is responsible for a large number of human cancers, not all smokers develop cancer. Similarly, some people who inherit mutations in the tumor-suppressor genes *TP53* or *RB1* never develop cancer. Explain these observations.

- **Hint:** *This problem asks you to consider the reasons why only some people develop cancer as a result of environmental factors or mutations in tumor-suppressor genes. The key to its solution is to consider the steps involved in the development of cancer and the number of abnormal functions in cancer cells. Also, consider how genetics may affect DNA repair functions.*

24.7 Viruses Contribute to Cancer in Both Humans and Animals

Viruses that cause cancer in animals have played a significant role in the search for knowledge about the genetics of human cancer. Most cancer-causing animal viruses are RNA viruses known as **retroviruses**. In humans, most of the known cancer-causing viruses are DNA viruses.

To understand how retroviruses cause cancer in animals, it is necessary to know how these viruses replicate in cells. When a retrovirus infects a cell, its RNA genome is copied into DNA by the **reverse transcriptase** enzyme, which is brought into the cell with the infecting virus. The DNA copy then enters the nucleus of the infected cell, where it integrates at random into the host cell's genome. The integrated DNA copy of the retroviral RNA is called a **provirus**. The proviral DNA contains powerful enhancer and promoter elements in its U5 and U3 sequences at the ends of the provirus (**Figure 24.11**). The proviral promoter uses the host cell's transcription proteins, directing transcription of the viral genes (*gag, pol,* and *env*). The products of these genes are the proteins and RNA genomes that make up the new retroviral particles. Because the provirus is integrated into the host genome, it is replicated along with the host's DNA during the cell's normal cell cycle. A retrovirus may not kill a cell, but it may continue to use the cell as a factory to replicate more viruses that will then infect surrounding cells.

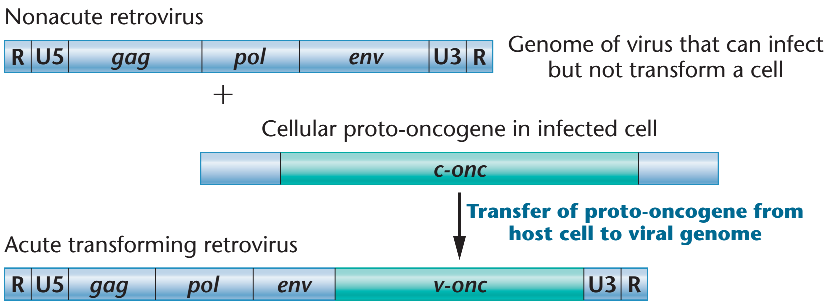

Nonacute retrovirus
R
U5
gag
pol
env
U3
R
Genome of virus that can infect but not transform a cell
+
Cellular proto-oncogene in infected cell
c-onc
Transfer of proto-oncogene from host cell to viral genome
Acute transforming retrovirus
R
U5
gag
pol
env
v-onc
U3
R

FIGURE 24.11 The genome of a typical retrovirus is shown at the top of the diagram. The genome contains repeats at the termini (R), the U5 and U3 regions that contain promoter and enhancer elements, and the three major genes that encode viral proteins (*gag, env,* and the viral reverse transcriptase *pol*). RNA transcripts of the entire viral genome comprise the new viral genomes. If the retrovirus acquires all or part of a host-cell proto-oncogene (*c-onc*), this gene (now known as a *v-onc*) is expressed along with the viral genes, leading to overexpression or inappropriate expression of the *v-onc* gene. The *v-onc* gene may also acquire mutations that enhance its transforming ability.

A retrovirus may cause cancer in three different ways. First, the proviral DNA may integrate by chance near one of the cell's normal proto-oncogenes. The strong promoters and enhancers in the provirus then stimulate high levels or inappropriate timing of transcription of the proto-oncogene, leading to stimulation of host-cell proliferation. Second, a retrovirus may pick up a copy of a host proto-oncogene and integrate it into its genome (Figure 24.11). The cellular proto-oncogene may be mutated during the process of transfer into the virus, or it may be expressed at abnormal levels because it is now under the control of viral promoters. Retroviruses that carry these cell-derived oncogenes can infect and transform normal cells into tumor cells and are known as **acute transforming retroviruses**. Through the study of many acute transforming viruses of animals, scientists have identified dozens of proto-oncogenes. Third, a retrovirus may contain a normal viral gene whose product can either stimulate the cell cycle or act as a gene-expression regulator for both cellular and viral genes. As a result, expression of such a viral gene may lead to inappropriate cell growth or to abnormal expression of cancer-related cellular genes.

So far, no acute transforming retroviruses have been identified in humans; however, several human retroviruses, such as human immunodeficiency virus (HIV) and the human T-cell leukemia virus (HTLV-1), are associated with human cancers. These retroviruses are thought to stimulate cancer development through the third mechanism, described in the previous paragraph.

DNA viruses also contribute to the development of human cancers in a variety of ways. Because viruses are composed solely of a nucleic acid genome surrounded by a protein coat, they must utilize the host cell's biosynthetic machinery to reproduce themselves. To access the host's DNA-synthesizing enzymes, viruses require the host cell to be in an actively growing state. Thus, many DNA viruses contain genes encoding products that stimulate the cell cycle. These products often interact with tumor-suppressor proteins, inactivating them. If the host cell survives the infection, it may lose control of the cell cycle and begin its journey to carcinogenesis.

It is thought that, worldwide, about 12 percent of human cancers are associated with viruses, making virus infection the second greatest risk factor for cancer, next to tobacco smoking. The most significant contributors to virus-induced cancers are listed in **Table 24.3**. Like other risk factors for cancer, including hereditary predisposition to certain cancers, virus infection alone is not sufficient to trigger human cancers. Other factors, including DNA damage or the accumulation of mutations in one or more of a cell's oncogenes and tumor-suppressor genes, are required to move a cell down the multistep pathway to cancer.

TABLE 24.3 Human Viruses Associated with Cancers

Virus		Associated Cancers
DNA Viruses		
Epstein-Barr virus	EBV	Burkitt lymphoma, nasopharyngeal carcinoma, Hodgkin lymphoma
Hepatitis B virus	HBV	Hepatocellular carcinoma
Hepatitis C virus	HCV	Hepatocellular carcinoma, non-Hodgkin lymphoma
Human papilloma viruses 16, 18	HPV16, 18	Cervical cancer, anogenital cancers, oral cancers
Kaposi sarcoma-associated herpesvirus	KSHV	Kaposi sarcoma, primary effusion lymphoma
Retroviruses		
Human T-cell lymphotropic virus, type 1	HTLV-1	Adult T-cell leukemia and lymphoma
Human immunodeficiency virus, type 1	HIV-1	Immune suppression, leading to cancers caused by other viruses (KSHV, EBV, HPV)

24.8 Environmental Agents Contribute to Human Cancers

Any substance or process that damages DNA has the potential to be carcinogenic. Unrepaired or inaccurately repaired DNA introduces mutations, which, if they occur in proto-oncogenes or tumor-suppressor genes, can lead to abnormal regulation of cell proliferation or disruption of controls over apoptosis or metastasis.

Our environment, both natural and human-made, contains abundant carcinogens. These include chemicals, radiation, some viruses, and chronic infections. In this section, we will examine some of these environmental agents that contribute to the development of cancer.

Natural Environmental Agents

Although most people perceive human-made, industrial chemicals to be the most significant contributors to cancer, they may contribute to less than 20 percent of cancers. Some of the most mutagenic agents, and hence potentially the most carcinogenic, are natural substances and natural processes. For example, aflatoxin, a component of a mold that grows on peanuts and corn, is one of the most carcinogenic chemicals known. Most chemical carcinogens, such as nitrosamines, are components of synthetic substances and are found in some preserved meats; however, many are naturally occurring. For example, natural pesticides and antibiotics found in plants may be carcinogenic, and the human body itself creates alkylating agents in the acidic environment of the gut.

DNA lesions brought about by natural radiation, metabolism, and DNA replication contribute significantly to the development of cancer. Normal metabolism creates oxidative end products that can damage DNA, proteins, and lipids. It is estimated that the human body suffers about 10,000 damaging DNA lesions per day due to the actions of oxygen free radicals. DNA repair enzymes deal successfully with most of this damage; however, some damage may lead to permanent mutations. The process of DNA replication itself is mutagenic. Hence, substances such as growth factors or hormones that stimulate cell division are ultimately mutagenic and perhaps carcinogenic. Chronic inflammation due to infection also stimulates tissue repair and cell division, resulting in DNA lesions accumulating during replication. These mutations may persist, particularly if cell-cycle checkpoints are compromised due to mutations or inactivation of tumor-suppressor genes such as *TP53* or *RB1*.

As we learned in Chapter 15, both ultraviolet (UV) light and ionizing radiation (such as X rays and gamma rays) induce DNA damage. The UV radiation in sunlight is well accepted as an inducer of skin cancers. Ionizing radiation has clearly shown itself to be a carcinogen in studies of populations exposed to neutron and gamma radiation from atomic blasts such as those in Hiroshima and Nagasaki. Another significant environmental component, radon gas, may be responsible for about 50 percent of the ionizing radiation exposure of the U.S. population and could contribute to lung cancers in some populations.

Diet is often implicated in the development of cancer. It is estimated that approximately 20 percent of cancer deaths in the United States are linked to dietary influences in conjunction with obesity and physical inactivity. Consumption of red meat and animal fat is associated with some cancers, such as colon, prostate, and breast cancer. The mechanisms by which these substances may contribute to carcinogenesis may involve stimulation of cell division through hormones or creation of carcinogenic chemicals during cooking, processing, or digestion. Alcohol may cause inflammation of the liver and contribute to liver cancer. It is also linked to other cancers including breast, colon, and esophagus.

Human-Made Chemicals and Pollutants

Although lifestyle factors such as smoking make significant contributions to the development of cancers, exposure to human-made carcinogens in air, food, and water also contribute. For example, researchers estimate that approximately 15 percent of all lung cancer deaths are due to components of air pollution, such as particulate matter. The International Agency for Research on Cancer (IARC) has tested more than 900 natural and artificial chemicals and found that more than 400 of them show some degree of carcinogenic properties in laboratory or epidemiological studies. Despite these test results, only a small fraction of the approximately 80,000 industrial chemicals in use today have been tested for carcinogenicity.

At least two problems make it difficult to estimate the effects of human-made chemicals and pollutants on human cancer. First, even those chemicals that have been found to be carcinogenic in laboratory and animal tests may not have detectable effects in humans, as the dosages are much lower in the environment than in the lab. Second, some carcinogens may not show their effects unless found in low-level mixtures with other toxic materials or when exposures occur in certain susceptible subpopulations such as infants or pregnant women.

Epidemiologic studies in humans contribute some information, but, because humans do not live in controlled environments, it is difficult to isolate the effects of one chemical agent when humans are exposed to hundreds of environmental chemicals, in varying dosages, and have differing genetic backgrounds.

The IARC has been testing carcinogens for several decades and has classified approximately 100 chemical agents (natural and human-made) as "carcinogenic to humans." Other candidate carcinogens can only be classified as probable, possible, or unknown carcinogens.

Tobacco Smoke and Cancer

One of the most thoroughly studied environmental and lifestyle carcinogens is tobacco smoke. Tobacco smoking is associated with at least 17 different types of human cancer, including lung cancers and cancers of the oral cavity, bladder, liver, stomach and kidney. It is estimated that tobacco smoking kills more than six million people each year, worldwide. Seventy percent of lung cancer deaths and more than 25 percent of all cancers can be linked to tobacco smoking.

Tobacco smoke contains a mixture of more than 4000 chemicals, and more than 60 of these are carcinogens. Well-known examples of these include benzene, arsenic, benzo[a]pyrene, cadmium, formaldehyde, and styrene. In addition, the particulate matter in smoke is a carcinogen.

Tobacco smoking triggers a large number of somatic mutations and epigenetic changes. Smoking one pack of cigarettes each day can create more than 150 mutations per year in the genomes of lung cells, as well as dozens of mutations in cells of the larynx, mouth, bladder, and liver. These mutations include base substitutions, insertions, deletions, copy-number aberrations in parts of chromosomes, and covalent bonding of reactive chemical adducts to DNA bases. These types of DNA damage can occur anywhere in the genome and may create driver mutations in proto-oncogenes or tumor-suppressor genes.

Lung cancer genomes from smokers also contain changes in DNA methylation patterns. Approximately 0.1 percent of CpG sequences that have been examined are either hypermethylated or hypomethylated. These changes, if present in regulatory regions of proto-oncogenes or tumor-suppressor genes, may contribute to altered gene expression in these cancers.

According to the World Health Organization and the American Cancer Society, smokers who quit tobacco smoking cut their risk of developing lung and other cancers by one-half within 5 years of quitting. They also see a reduced risk of developing coronary heart disease and diabetes.

Now Solve This

24.4 Cancer can arise spontaneously, but it can also be induced as a result of environmental factors such as sun exposure, infections, and tobacco smoking. If you were asked to help allocate resources to cancer research, what emphasis would you place on research to find cancer cures, compared to that placed on education about cancer prevention?

■ **Hint:** *This problem asks you to consider the outcomes of two different approaches to cancer research. The key to its solution is to think about the relative rates of environmentally induced and spontaneous cancers. [An interesting source of information on this topic is Ames, B. N. et al. (1995). The causes and prevention of cancer.* Proc. Natl. Acad. Sci. USA *92:5258–5265.)*

CASE STUDY Cancer-killing bacteria

Ralph, a 57-year-old man, was diagnosed with colon cancer. His oncologist discussed the use of radiation and chemotherapy as treatments, both of which can cause debilitating side effects. Ralph decided to explore other options and went to a cancer clinic. He learned that researchers in a synthetic biology program were testing the use of genetically modified bacterial cells designed to selectively invade specific tumors and kill cancer cells, with no effects on normal cells. Ralph decided to participate and was informed that he would be part of a phase III trial, comparing the effects of the modified bacterial cell treatment against conventional chemotherapy. However, as part of the trial, he would be randomly assigned to receive one or the other treatment. He was disappointed to learn this, because he assumed that he would receive the bacterial therapy.

1. Informed consent is legally and ethically required before someone participates in a clinical trial. After potential participants receive information about the trial and what constitutes informed consent, research indicates that 25 percent of prospective participants do not understand that these trials are designed primarily to establish the efficacy of the treatment rather than directly benefit participants. What should investigators do to make sure that clinical trial participants understand that the trial is not primarily intended to help them?

2. If you were in Ralph's position, would you try radiation and chemotherapy instead, or enroll in the trial on the chance that you might receive the bacterial therapy, which may or may not be more effective than the conventional therapy?

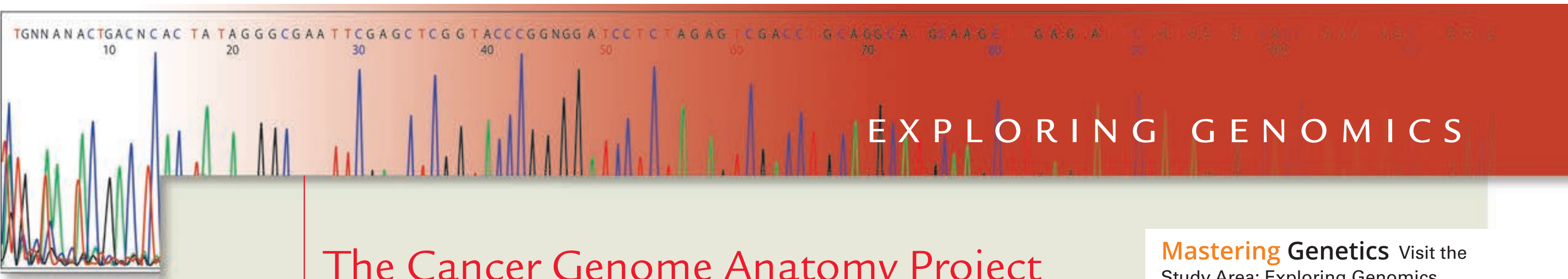
3. If you agree to participate and then learn that you will not be receiving the bacterial treatment, would you be ethically bound to continue in the trial?

See Joffe , S., et al. (2001). Quality of informed consent in cancer clinical trials: a cross-sectional survey. *Lancet* 358(9295):1772–1777.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Cancer cells show two fundamental properties: abnormal cell proliferation and a propensity to spread and invade other parts of the body.
2. Cancers are clonal, meaning that all cells within a tumor originate from a single cell that contained a number of driver mutations.
3. The development of cancer is a multistep process, requiring mutations in several cancer-related genes.
4. Cancer cells contain gene mutations, chromosomal abnormalities, genomic instability, and abnormal patterns of chromatin modifications.
5. Cancer cells have defects in DNA damage repair, chromatin modifications, cell-cycle regulation, and programmed cell death.
6. Proto-oncogenes are normal genes that promote cell growth and division. When proto-oncogenes are mutated or misexpressed in cancer cells, they are known as oncogenes.
7. Tumor-suppressor genes normally regulate cell-cycle checkpoints and apoptosis. When tumor-suppressor genes are mutated or inactivated, cells cannot correct DNA damage. This leads to accumulations of mutations that may cause cancer.
8. The ability of cancer cells to metastasize requires gene products that control a number of functions such as cell adhesion, proteolysis, and tissue invasion.
9. Inherited mutations in cancer-susceptibility genes are not sufficient to trigger cancer. Other somatic mutations in proto-oncogenes or tumor-suppressor genes are necessary for the development of hereditary cancers.
10. Tumor viruses contribute to cancers by introducing viral oncogenes, interfering with tumor-suppressor proteins, or altering expression of a cell's proto-oncogenes.
11. Natural and human-made environmental agents such as chemicals, radiation, viruses, and chronic infections contribute to the development of cancer.



EXPLORING GENOMICS

The Cancer Genome Anatomy Project (CGAP)

Mastering Genetics Visit the Study Area: Exploring Genomics

A research group headed by Dr. Victor Velculescu of Johns Hopkins University reported that breast and colon cancers contain about 11 gene mutations that may contribute to the cancer phenotype. The research group analyzed 13,023 of the 21,000 known genes in the human genome, comparing the DNA sequences from normal cells and cancer cells. Most of the mutations that were specific to cancer cells were not previously known to be associated with cancer.

Dr. Velculescu's study was one of the first in *The Cancer Genome Atlas (TCGA)* project, a $1.5 billion federal project designed to systematically scan the human genome to find genes that are mutated in many different cancers. In this exercise, we will explore aspects of Dr. Velculescu's research by mining information available in the online database **Cancer Genome Anatomy Project (CGAP)**. The purpose of the CGAP is to understand the expression profiles of genes from normal, precancer, and cancer cells.

■ Exercise – Colon Cancer and the *TBX22* Gene

One gene that Dr. Velculescu's research group discovered to be mutated in colon cancers—*TBX22*—was not previously suspected to contribute to this cancer. What is *TBX22*, and how do you think a mutated *TBX22* gene would contribute to the development of colon cancer?

1. To begin your search for the answers, go to CGAP at https://cgap.nci.nih.gov.
2. Click the "Genes" button near the top of the page.
3. From the list of "Gene Tools" in the left-hand margin, select "Gene Finder."
4. Select "Homo sapiens" in the "Select organism" box, and type TBX22 in the "Enter a unique identifier" box. Submit the query.
5. Select "Gene Info" in the right-hand column of the table.
6. Explore the many sources of information about *TBX22* from various database links listed on this page.

Prepare a brief written or verbal report on what you learned during your explorations and which sources you used to reach your conclusions about *TBX22*.

(continued)

INSIGHTS AND SOLUTIONS

1. In retinoblastoma, a mutation in one allele of the *RB1* tumor-suppressor gene can be inherited from the germ line, causing an autosomal dominant predisposition to the development of eye tumors. To develop tumors, a somatic mutation in the second copy of the *RB1* gene is necessary, indicating that the mutation itself acts as a recessive trait. Given that the first mutation can be inherited, in what ways can a second mutational event occur?

   Solution: In considering how this second mutation arises, we must look at several types of mutational events, including changes in nucleotide sequence and events that involve whole chromosomes or chromosome parts. Retinoblastoma results when both copies of the *RB1* locus are lost or inactivated. With this in mind, you must first list the phenomena that can result in a mutational loss or the inactivation of a gene.

   One way the second *RB1* mutation can occur is by a nucleotide alteration that converts the remaining normal *RB1* allele to a mutant form. This alteration can occur through a nucleotide substitution or through a frameshift mutation caused by the insertion or deletion of nucleotides during replication. A second mechanism involves the loss of the chromosome carrying the normal allele. This event would take place during mitosis, resulting in chromosome 13 monosomy and leaving the mutant copy of the gene as the only *RB1* allele. This mechanism does not necessarily involve loss of the entire chromosome; deletion of the long arm (*RB1* is on 13q) or an interstitial deletion involving the *RB1* locus and some surrounding material would have the same result. Alternatively, a chromosome aberration involving loss of the normal copy of the *RB1* gene might be followed by duplication of the chromosome carrying the mutant allele. Two copies of chromosome 13 would be restored to the cell, but the normal *RB1* allele would not be present. Finally, a recombination event followed by chromosome segregation could produce a homozygous combination of mutant *RB1* alleles.

2. Proto-oncogenes can be converted to oncogenes in a number of different ways. In some cases, the proto-oncogene itself becomes amplified up to hundreds of times in a cancer cell. An example is the *cyclin D1* gene, which is amplified in some cancers. In other cases, the proto-oncogene may be mutated in a limited number of specific ways, leading to alterations in the gene product's structure. The *ras* gene is an example of a proto-oncogene that becomes oncogenic after suffering point mutations in specific regions of the gene. Explain why these two proto-oncogenes (*cyclin D1* and *ras*) undergo such different alterations to convert them into oncogenes.

   Solution: The first step in solving this question is to understand the normal functions of these proto-oncogenes and to think about how either amplification or mutation would affect each of these functions.

   The cyclin D1 protein regulates progression of the cell cycle from G1 into S phase, by binding to CDK4 and activating this kinase. The cyclin D1/CDK4 complex phosphorylates a number of proteins including pRB, which in turn activate other proteins in a cascade that results in transcription of genes whose products are necessary for DNA replication in S phase. The simplest way to increase the activity of cyclin D1 would be to increase the number of cyclin D1 molecules available for binding to the cell's endogenous CDK4 molecules. This can be accomplished by several mechanisms, including amplification of the *cyclin D1* gene. In contrast, a point mutation in the *cyclin D1* gene would most likely interfere with the ability of the cyclin D1 protein to bind to CDK4; hence, mutations within the gene would probably repress cell-cycle progression rather than stimulate it.

   The *ras* gene product is a signal transduction protein that operates as an on/off switch in response to external stimulation by growth factors. It does so by binding either GTP (the "on" state) or GDP (the "off" state). Oncogenic mutations in the *ras* gene occur in specific regions that alter the ability of the Ras protein to exchange GDP for GTP. Oncogenic Ras proteins are locked in the "on" conformation, bound to GTP. In this way, they constantly stimulate the cell to divide. An amplification of the *ras* gene would simply provide more molecules of normal Ras protein, which would still be capable of on/off regulation. Hence, simple amplification of *ras* would less likely be oncogenic.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on cancer as a genetic disease, with an emphasis on the relationship between cancer, the cell cycle, and DNA damage, as well as on the multiple steps that lead to cancer. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter,

   (a) How do we know that malignant tumors arise from a single cell that contains mutations?

   (b) How do we know that cancer development requires more than one mutation?

   (c) How do we know that cancer cells contain defects in DNA repair?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 579. These concepts relate to the multiple ways in which genetic alterations lead to the development of cancers. The sixth concept states that epigenetic effects including DNA methylation and histone modifications contribute to the genetic alterations

leading to cancer. Write a short essay describing how epigenetic changes in cancer cells contribute to the development of cancers.

3. Where are the major regulatory points in the cell cycle?

4. List the functions of kinases and cyclins, and describe how they interact to cause cells to move through the cell cycle.

5. How can mutations in noncoding segments of DNA contribute to the development of cancers?

6. What is the difference between saying that cancer is inherited and saying that the predisposition to cancer is inherited?

7. As a genetic counselor, you are asked to assess the risk for a couple with a family history of familial adenomatous polyposis (FAP) who are thinking about having children. Neither the husband nor the wife has colorectal cancer, but the husband has a sister with FAP. What is the probability that this couple will have a child with FAP? Are there any tests that you could recommend to help in this assessment?

8. What is apoptosis, and under what circumstances do cells undergo this process?

9. Define tumor-suppressor genes. Why is a mutated single copy of a tumor-suppressor gene expected to behave as a recessive gene?

10. Describe the steps by which the *TP53* gene responds to DNA damage and/or cellular stress to promote cell-cycle arrest and apoptosis. Given that *TP53* is a recessive gene and is not located on the X chromosome, why would people who inherit just one mutant copy of a recessive tumor-suppressor gene be at higher risk of developing cancer than those without the recessive gene?

11. Part of the Ras protein is associated with the plasma membrane, and part extends into the cytoplasm. How does the Ras protein transmit a signal from outside the cell into the cytoplasm? What happens in cases where the *ras* gene is mutated?

12. If a cell suffers damage to its DNA while in S phase, how can this damage be repaired before the cell enters mitosis?

13. Distinguish between oncogenes and proto-oncogenes. In what ways can proto-oncogenes be converted to oncogenes?

14. Of the two classes of genes associated with cancer, tumor-suppressor genes and oncogenes, mutations in which group can be considered gain-of-function mutations? In which group are the loss-of-function mutations? Explain.

15. How do translocations such as the Philadelphia chromosome contribute to cancer?

16. Explain why many oncogenic viruses contain genes whose products interact with tumor-suppressor proteins.

17. DNA sequencing has provided data to indicate that cancer cells may contain tens of thousands of somatic mutations, only some of which confer a growth advantage to a cancer cell. How do scientists describe and categorize these recently discovered populations of mutations in cancer cells?

18. How do normal cells protect themselves from accumulating mutations in genes that could lead to cancer? How do cancer cells differ from normal cells in these processes?

19. Describe the difference between an acute transforming virus and a virus that does not cause tumors.

20. Epigenetics is a relatively new area of genetics with a focus on phenomena that affect gene expression but do not affect DNA sequence. Epigenetic effects are quasi-stable and may be passed to progeny somatic or germ-line cells. What are known causes of epigenetic effects, and how do they relate to cancer?

21. Radiotherapy (treatment with ionizing radiation) is one of the most effective current cancer treatments. It works by damaging DNA and other cellular components. In which ways could radiotherapy control or cure cancer, and why does radiotherapy often have significant side effects?

22. Genetic tests that detect mutations in the *BRCA1* and *BRCA2* tumor-suppressor genes are widely available. These tests reveal a number of mutations in these genes—mutations that have been linked to familial breast cancer. Assume that a young woman in a suspected breast cancer family takes the *BRCA1* and *BRCA2* genetic tests and receives negative results. That is, she does not test positive for the mutant alleles of *BRCA1* or *BRCA2*. Can she consider herself free of risk for breast cancer?

23. Explain the apparent paradox that both hypermethylation and hypomethylation of DNA are often found in the same cancer cell.

24. As part of a cancer research project, you have discovered a gene that is mutated in many metastatic tumors. After determining the DNA sequence of this gene, you compare the sequence with those of other genes in the human genome sequence database. Your gene appears to code for an amino acid sequence that resembles sequences found in some serine proteases. Conjecture how your new gene might contribute to the development of highly invasive cancers.

Extra-Spicy Problems

25. Mutations in tumor-suppressor genes are associated with many types of cancers. In addition, epigenetic changes (such as DNA methylation) of tumor-suppressor genes are also associated with tumorigenesis [Otani et al. (2013). *Expert Rev Mol Diagn* 13:445–455].
(a) How might hypermethylation of the *TP53* gene promoter influence tumorigenesis?
(b) Knowing that tumors release free DNA into certain surrounding body fluids through necrosis and apoptosis Kloten et al. [(2013). *Breast Cancer Res*. 15(1):R4] outline an experimental protocol for using human blood as a biomarker for cancer and as a method for monitoring the progression of cancer in an individual.

26. A study by Bose and colleagues [(1998). *Blood* 92:3362–3367] and a previous study by Biernaux and others [(1996). *Bone Marrow Transplant* 17:(Suppl. 3) S45–S47] showed that *BCR-ABL* fusion gene transcripts can be detected in 25 to 30 percent of healthy adults who do not develop chronic myelogenous leukemia (CML). Explain how these individuals can carry a fusion gene that is transcriptionally active and yet do not develop CML.

27. Those who inherit a mutant allele of the *RB1* tumor-suppressor gene are at risk for developing a bone cancer called osteosarcoma. You suspect that in these cases, osteosarcoma requires a mutation in the second *RB1* allele, and you have cultured some osteosarcoma cells and obtained a cDNA clone of

a normal human *RB1* gene. A colleague sends you a research paper revealing that a strain of cancer-prone mice develop malignant tumors when injected with osteosarcoma cells, and you obtain these mice. Using these three resources, what experiments would you perform to determine (a) whether osteosarcoma cells carry two *RB1* mutations, (b) whether osteosarcoma cells produce any pRB protein, and (c) if the addition of a normal *RB1* gene will change the cancer-causing potential of osteosarcoma cells?

28. The table in this problem summarizes some of the data that have been collected on mutations in the *BRCA1* tumor-suppressor gene in families with a high incidence of both early-onset breast cancer and ovarian cancer.

Predisposing Mutations in *BRCA1*

Kindred	Codon	Nucleotide Change	Coding Effect	Frequency in Control Chromosomes
1901	24	−11 bp	Frameshift or splice	0/180
2082	1313	C → T	Gln → Stop	0/170
1910	1756	Extra C	Frameshift	0/162
2099	1775	T → G	Met → Arg	0/120
2035	NA*	?	Loss of transcript	NA*

Source: (1994). *Science* 266:66–71. © AAAS.

*NA indicates not applicable, as the regulatory mutation is inferred and the position has not been identified.

(a) Note the coding effect of the mutation found in kindred group 2082. This results from a single base-pair substitution. Draw the normal double-stranded DNA sequence for this codon (with the 5′ and 3′ ends labeled), and show the sequence of events that generated this mutation, assuming that it resulted from an uncorrected mismatch event during DNA replication.

(b) Examine the types of mutations that are listed in the table, and determine if the *BRCA1* gene is likely to be a tumor-suppressor gene or an oncogene.

(c) Although the mutations listed in the table are clearly deleterious and cause breast cancer in women at very young ages, each of the kindred groups had at least one woman who carried the mutation but lived until age 80 without developing cancer. Name at least two different mechanisms (or variables) that could underlie variation in the expression of a mutant phenotype, and propose an explanation for the incomplete penetrance of this mutation. How do these mechanisms or variables relate to this explanation?

29. Researchers have identified some tumors that have no recurrent mutations or deletions in known oncogenes or tumor-suppressor genes and no detectable epigenetic alterations. However, these tumors often have large chromosomal deletions. What are some possible explanations that could account for the genetic causes behind these tumors?

30. Although cancer is not a contagious disease in humans or other vertebrates, there have been rare cases in which cancers have spread from one organism to another. Describe three cases of these contagious cancers and what conditions might have led to their appearance. For an introduction to this topic, see http://www.cancer.org/cancer/cancerbasics/is-cancer-contagious.

25

Quantitative Genetics and Multifactorial Traits

A field of pumpkins, where size is under the influence of additive alleles.

CHAPTER CONCEPTS

- Quantitative inheritance results in a range of measurable phenotypes for a polygenic trait.
- With some exceptions, polygenic traits tend to demonstrate continuous variation.
- Quantitative traits can be explained in Mendelian terms whereby certain alleles have an additive effect on the traits under study.
- The study of polygenic traits relies on statistical analysis.
- Heritability values estimate the genetic contribution to phenotypic variability under specific environmental conditions.
- Twin studies allow an estimation of heritability in humans.
- Quantitative trait loci (QTLs) can be mapped and identified.

Up to this point in the text, most of our examples of phenotypic variation have been those that have been assigned to distinct and separate categories; for example, human blood type was A, B, AB, or O; squash fruit shape was spherical, disc shaped, or elongated; and fruit fly eye color was red or white (see Chapter 4). Typically with these traits, a genotype will produce a single identifiable phenotype, although phenomena such as variable penetrance and expressivity, pleiotropy, and epistasis can obscure the relationship between genotype and phenotype.

In this chapter, we will look at traits that are not as clear cut, including many that are of medical or agricultural importance. The traits on which we focus show much more variation, often falling into a continuous range of phenotypes that are more difficult to classify into distinct categories. Most of these traits show *continuous variation*, including, for example, height in humans, milk and meat production in cattle, and yield and seed protein content in various crops. Continuous variation across a range of phenotypes is measured and described in quantitative terms, so this genetic phenomenon is known as **quantitative inheritance**. And because the varying phenotypes result from the input of genes at more than one, and often many, loci, they are also said to be **polygenic** (literally "of many genes").

To further complicate the link between the genotype and phenotype, the genotype generated at fertilization establishes a quantitative range within which a particular individual can fall. However, the final phenotype is often also influenced by environmental factors to which that individual is exposed. Human height, for example, is genetically influenced but is also affected by environmental factors such as nutrition. Quantitative (polygenic) traits whose phenotypes result from both gene action and environmental influences are often termed **multifactorial**, or **complex traits**. Often these terms are used interchangeably. For consistency throughout the chapter, we will utilize the term *multifactorial* in our discussions.

In this chapter, we will examine examples of quantitative inheritance, multifactorial traits, and some of the statistical techniques used to study them. We will also consider how geneticists assess the relative importance of genetic versus environmental factors contributing to continuous phenotypic variation, and we will discuss approaches to identifying and mapping genes that influence quantitative traits.

25.1 Not All Polygenic Traits Show Continuous Variation

In addition to quantitative traits that display continuous variation, there are two other classes of polygenic traits. **Meristic traits** are those in which the phenotypes are described by whole numbers. Examples of meristic traits include the number of seeds in a pod or the number of eggs laid by a chicken in a year. These are quantitative traits, but they do not have an infinite range of phenotypes: for example, a pod may contain 2, 4, or 6 seeds, but not 5.75. **Threshold traits** are polygenic (and frequently multifactorial), but they are distinguished from continuous and meristic traits by having a small number of discrete phenotypic classes. Threshold traits are currently of heightened interest to human geneticists because an increasing number of diseases are now thought to show this pattern of polygenic inheritance. One example is **Type II diabetes**, also known as adult-onset diabetes because it typically affects individuals who are middle aged or older. A population can be divided into just two phenotypic classes for this trait—individuals who have Type II diabetes and those who do not—so at first glance, this human disease may appear to more closely resemble a simple monogenic trait. However, no single adult-onset diabetes gene has been identified. Instead, the combination of alleles present at multiple contributing loci gives an individual a greater or lesser likelihood of developing the disease. These varying levels of liability form a continuous range: At one extreme are those at very low risk for Type II diabetes, while at the other end of the distribution are those whose genotypes make it highly likely they will develop the disease (Figure 25.1). As with many threshold traits, environmental factors also play a role in determining the final phenotype, with diet and lifestyle having a significant impact on whether an individual with moderate to high genetic liability will actually develop Type II diabetes.

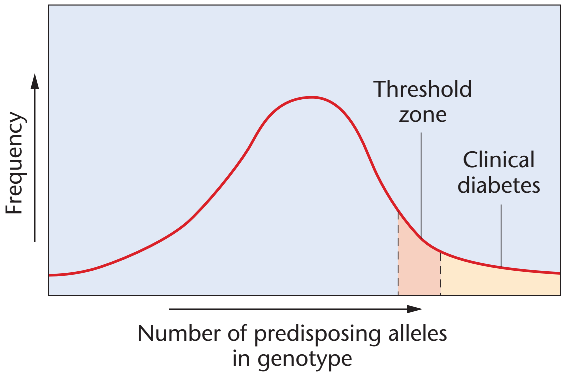


FIGURE 25.1 A graphic depiction of predisposing alleles characteristic of a threshold trait within a population, illustrated by Type II diabetes.

25.2 Quantitative Traits Can Be Explained in Mendelian Terms

The question of whether continuous phenotypic variation could be explained in Mendelian terms caused considerable controversy in the early 1900s. Some scientists argued that, although Mendel's unit factors, or genes, explained patterns of discontinuous segregation with discrete phenotypic classes, they could not also account for the range of phenotypes seen in quantitative patterns of inheritance. However, geneticists William Bateson and G. Udny Yule, adhering to a Mendelian explanation, proposed the **multiple-factor** or **multiple-gene hypothesis**, in which many genes, each individually behaving in a Mendelian fashion, contribute to the phenotype in a *cumulative* or *quantitative* way.

The Multiple-Gene Hypothesis for Quantitative Inheritance

The multiple-gene hypothesis was initially based on a key set of experimental results published by Herman Nilsson-Ehle in 1909. Nilsson-Ehle used grain color in wheat to test the concept that the cumulative effects of alleles at multiple loci produce the range of phenotypes seen in quantitative traits. In one set of experiments, wheat with red grain was crossed to wheat with white grain (Figure 25.2). The F_1 generation demonstrated an intermediate pink color, which at first sight suggested incomplete dominance of two alleles at a single locus. However, in the F_2 generation, Nilsson-Ehle did not observe the typical segregation of a monohybrid cross. Instead, approximately 15/16 of the plants showed some degree of red grain color, while 1/16 of the plants showed white grain color. Careful examination of the F_2 revealed that grain with color could be classified into four different shades of red. Because the F_2 ratio occurred in sixteenths, it appears that two genes, each with two alleles, control the phenotype and that they segregate independently from one another in a Mendelian fashion.

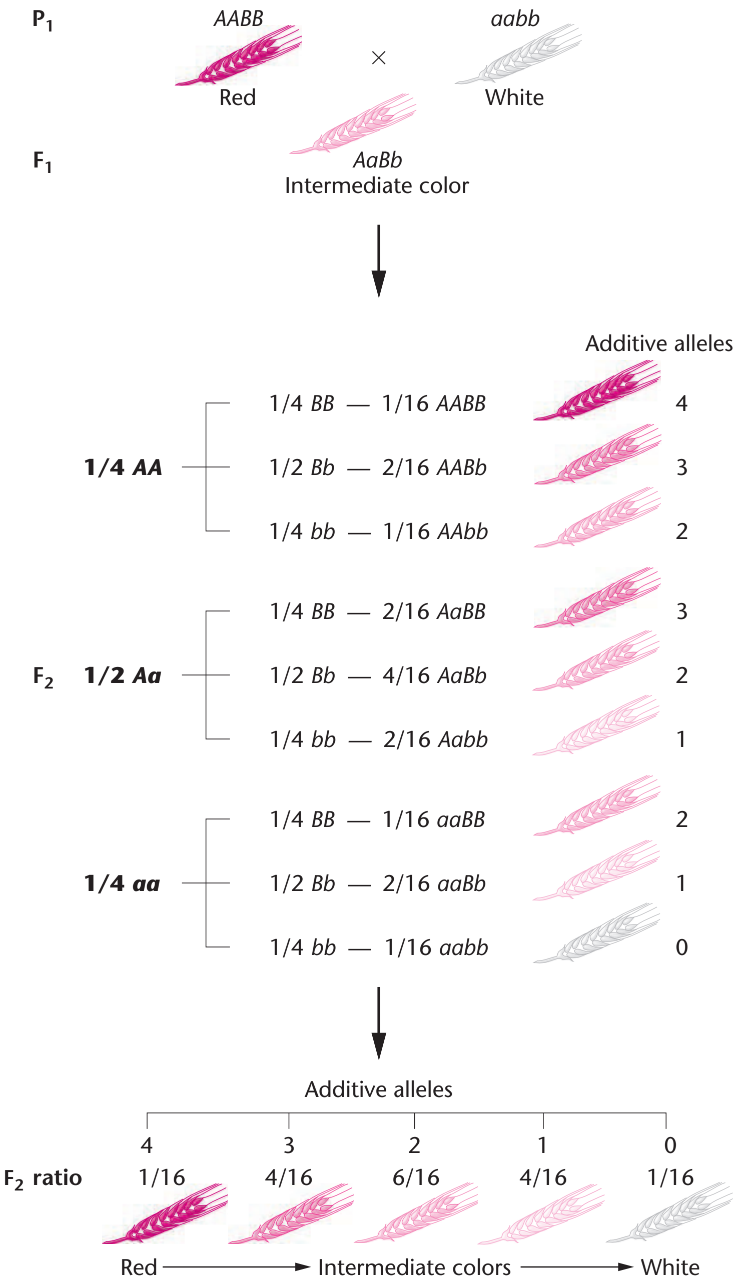


FIGURE 25.2 How the multiple-factor hypothesis accounts for the 1 : 4 : 6 : 4 : 1 phenotypic ratio of grain color when all alleles designated by an uppercase letter are additive and contribute an equal amount of pigment to the phenotype.

If each gene has one potential **additive allele** that contributes approximately equally to the red grain color and one potential **nonadditive allele** that fails to produce any red pigment, we can see how the multiple-factor hypothesis could account for the various grain color phenotypes. In the P_1 both parents are homozygous; the red parent contains only additive alleles (*AABB* in Figure 25.2), while the white parent contains only nonadditive alleles (*aabb*). The F_1 plants are heterozygous (*AaBb*), contain two additive (*A* and *B*) and two nonadditive (*a* and *b*) alleles, and express the intermediate pink phenotype. Each of the F_2 plants has 4, 3, 2, 1, or 0 additive alleles. F_2 plants with no additive alleles are white (*aabb*) like one of the P_1 parents, while F_2 plants with 4 additive alleles are red (*AABB*) like the other P_1 parent. Plants with 3, 2, or 1 additive alleles constitute the other three categories of red color observed in the F_2 generation. The greater the number of additive alleles in the genotype, the more intense the red color expressed in the phenotype, as each additive allele present contributes equally to the cumulative amount of pigment produced in the grain.

Nilsson-Ehle's results showed how continuous variation could still be explained in a Mendelian fashion, with additive alleles at multiple loci influencing the phenotype in a quantitative manner, but each individual allele segregating according to Mendelian rules. As we saw in Nilsson-Ehle's initial cross, if two loci, each with two alleles, were involved, then five F_2 phenotypic categories in a 1 : 4 : 6 : 4 : 1 ratio would be expected. However, there is no reason why three, four, or more loci cannot function in a similar fashion in controlling various quantitative phenotypes. As more quantitative loci become involved, greater and greater numbers of classes appear in the F_2 generation in more complex ratios. The number of phenotypes and the expected F_2 ratios for crosses involving up to five gene pairs are illustrated in Figure 25.3.

Additive Alleles: The Basis of Continuous Variation

The multiple-gene hypothesis consists of the following major points:

1. Phenotypic traits showing continuous variation can be quantified by measuring, weighing, counting, and so on.

2. Two or more gene loci, often scattered throughout the genome, account for the hereditary influence on the phenotype in an *additive way*. Because many genes may be involved, inheritance of this type is called *polygenic*.

3. Each gene locus may be occupied by either an *additive* allele, which contributes a constant amount to the phenotype, or a *nonadditive* allele, which does not contribute quantitatively to the phenotype.

4. The contribution to the phenotype of each additive allele, though often small, is approximately equal. While we now know this is not always true, we have made this assumption in the above discussion.

5. Together, the additive alleles contributing to a single quantitative character produce substantial phenotypic variation.

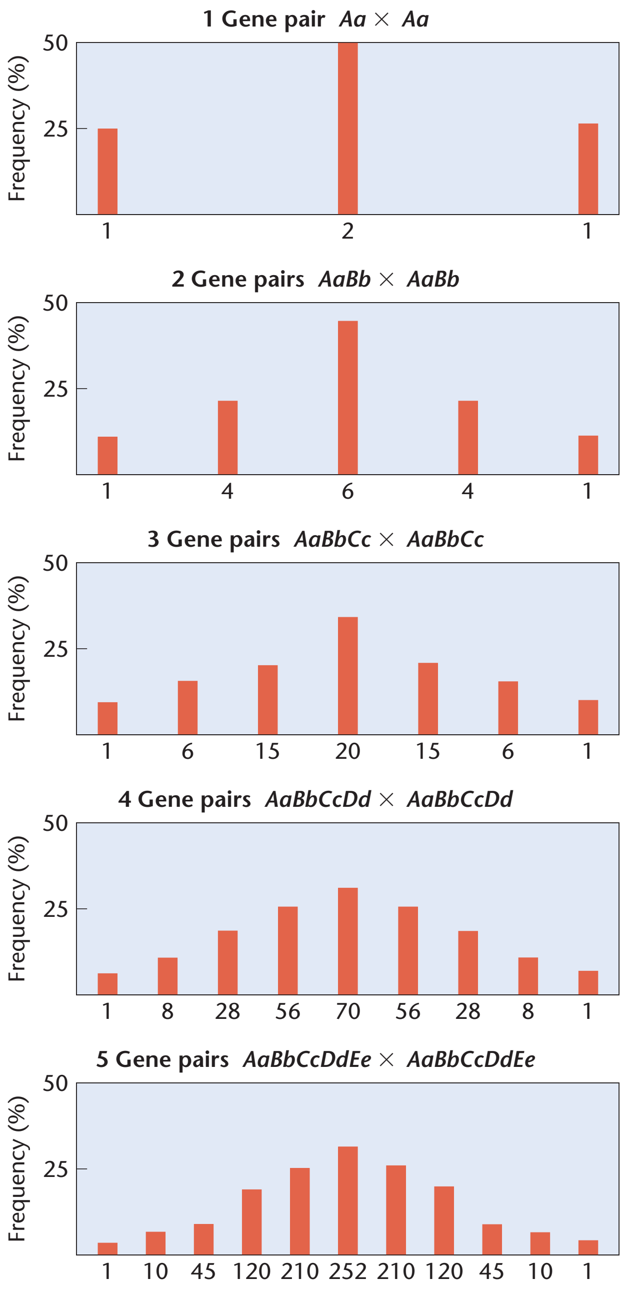


FIGURE 25.3 The genetic ratios (on the *x*-axis) resulting from crossing two heterozygotes when polygenic inheritance is in operation with 1–5 gene pairs. The histogram bars indicate the distinct F_2 phenotypic classes, ranging from one extreme (left end) to the other extreme (right end). Each phenotype results from a different number of additive alleles.

Calculating the Number of Polygenes

Various formulas have been developed for estimating the number of **polygenes**, the genes contributing to a quantitative trait. For example, if the ratio of F_2 individuals resembling *either* of the two extreme P_1 phenotypes can be determined, the number of polygenes involved (n) may be calculated as follows:

$$1/4^n = \text{ratio of } F_2 \text{ individuals expressing either extreme phenotype}$$

In the example of the red and white wheat grain color summarized in Figure 25.2, 1/16 of the progeny are either red *or* white like the P_1 phenotypes. This ratio can be substituted on the right side of the equation to solve for n:

$$\frac{1}{4^n} = \frac{1}{16}$$

$$\frac{1}{4^2} = \frac{1}{16}$$

$$n = 2$$

Table 25.1 lists the ratio and the number of F_2 phenotypic classes produced in crosses involving up to five gene pairs.

For low numbers of polygenes (n), it is sometimes easier to use the equation

$$(2n + 1) = \text{the number of distinct phenotypic categories observed}$$

For example, when there are two polygenes involved ($n = 2$), then $(2n + 1) = 5$ and each phenotype is the result of 4, 3, 2, 1, or 0 additive alleles. If $n = 3$, $2n + 1 = 7$ and each phenotype is the result of 6, 5, 4, 3, 2, 1, or 0 additive alleles. Thus, working backward with this rule and knowing the number of phenotypes, we can calculate the number of polygenes controlling them.

It should be noted, however, that both of these simple methods for estimating the number of polygenes involved in a quantitative trait assume not only that all the relevant alleles contribute equally and additively, but also that phenotypic expression in the F_2 is not affected significantly by environmental factors. As we will see later, for many quantitative traits, these assumptions may not be true.

NOW SOLVE THIS

25.1 A homozygous plant with 20-cm-diameter flowers is crossed with a homozygous plant of the same species that has 40-cm-diameter flowers. The F_1 plants all have flowers 30 cm in diameter. In the F_2 generation of 512 plants, 2 plants have flowers 20 cm in diameter, 2 plants have flowers 40 cm in diameter, and the remaining 508 plants have flowers of a range of sizes in between.

(a) Assuming all alleles involved act additively, how many genes control flower size in this plant?

(b) What frequency distribution of flower diameter would you expect to see in the progeny of a backcross between an F_1 plant and the large-flowered parent?

■ **HINT:** *This problem provides F_1 and F_2 data for a cross involving a quantitative trait and asks you to calculate the number of genes controlling the trait. The key to its solution is to remember that unless you know the total number of distinct F_2 phenotypes involved, the ratio (not the number) of parental phenotypes reappearing in the F_2 must be used in your determination of the number of genes involved.*

For more practice, see Problems 4–7.

TABLE 25.1 Determination of the Number of Polygenes (*n*) Involved in a Quantitative Trait

n	Individuals Expressing Either Extreme Phenotype	Distinct Phenotypic Classes
1	1/4	3
2	1/16	5
3	1/64	7
4	1/256	9
5	1/1024	11

25.3 The Study of Polygenic Traits Relies on Statistical Analysis

Before considering the approaches that geneticists use to dissect how much of the phenotypic variation observed in a population is due to genotypic differences among individuals and how much is due to environmental factors, we need to consider the basic statistical tools they use for the task. It is not usually feasible to measure expression of a polygenic trait in every individual in a population, so a random subset of individuals is usually selected for measurement to provide a *sample*. It is important to remember that the accuracy of the final results of the measurements depends on whether the sample is truly random and representative of the population from which it was drawn. Suppose, for example, that a student wants to determine the average height of the 100 students in his genetics class, and for his sample he measures the two students sitting next to him, both of whom happen to be centers on the college basketball team. It is unlikely that this sample will provide a good estimate of the average height of the class, for two reasons: First, it is too small; second, it is not a representative subset of the class (unless all 100 students are centers on the basketball team).

If the sample measured for expression of a quantitative trait is sufficiently large and it is also representative of the population from which it is drawn, we often find that the data form a **normal distribution**; that is, they produce a characteristic bell-shaped curve when plotted as a frequency histogram (**Figure 25.4**). Several statistical concepts are useful in the analysis of traits that exhibit a normal distribution, including the mean, variance, standard deviation, standard error of the mean, and covariance.

The Mean

The mean provides information about where the central point lies along a range of measurements for a quantitative trait. **Figure 25.5** shows the distribution curves for two different sets of phenotypic measurements. Each of these sets of measurements clusters around a central value (as it happens, they both cluster around the same value). This clustering is called a central tendency, and the central point is the mean.

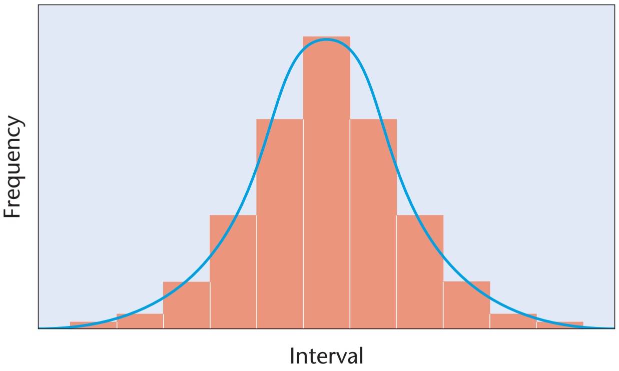


FIGURE 25.4 Normal frequency distribution, characterized by a bell-shaped curve.

Specifically, the **mean ($\overline{X}$)** is the arithmetic average of a set of measurements and is calculated as

$$\overline{X} = \frac{\Sigma X_i}{n}$$

where $\overline{X}$ is the mean, ΣX_i represents the sum of all individual values in the sample, and n is the number of individual values.

The mean provides a useful descriptive summary of the sample, but it tells us nothing about the range or spread of the data. As illustrated in Figure 25.5, a symmetrical distribution of values in the sample may, in one case, be clustered near the mean. Or a set of measurements may have the same mean but be distributed more widely around it. A second statistic, the variance, provides information about the spread of data around the mean.

Variance

The **variance (s^2)** for a sample is the average squared distance of all measurements from the mean. It is calculated as

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n - 1}$$

where the sum (Σ) of the squared differences between each measured value (X_i) and the mean ($\overline{X}$) is divided by one less than the total sample size $n - 1$.

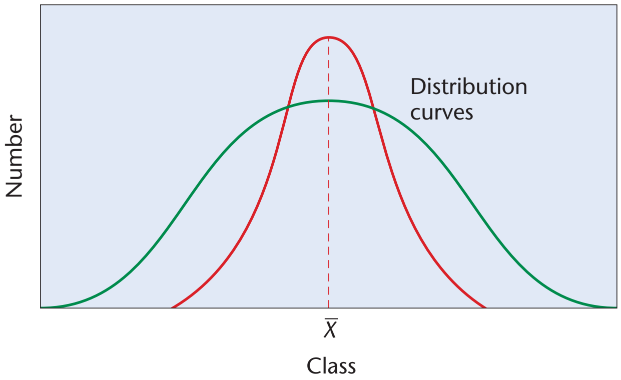


FIGURE 25.5 Two normal frequency distributions with the same mean but different amounts of variation.

As Figure 25.5 shows, it is possible for two sets of sample measurements for a quantitative trait to have the same mean but a different distribution of values around it. This range will be reflected in different variances. Estimation of variance can be useful in determining the degree of genetic control of traits when the immediate environment also influences the phenotype.

Standard Deviation

Because the variance is a squared value, its unit of measurement is also squared (m^2, g^2, etc.). To express variation around the mean in the original units of measurement, we can use the square root of the variance, a term called the **standard deviation (*s*)**:

$$s = \sqrt{s^2}$$

Table 25.2 shows the percentage of individual values within a normal distribution that fall within different multiples of the standard deviation. The values that fall within one standard deviation to either side of the mean represent 68 percent of all values in the sample. More than 95 percent of all values are found within two standard deviations to either side of the mean. This means that the standard deviation *s* can also be interpreted in the form of a probability. For example, a sample measurement picked at random has a 68 percent probability of falling within the range of one standard deviation.

Standard Error of the Mean

If multiple samples are taken from a population and measured for the same quantitative trait, we might find that their means vary. Theoretically, larger, truly random samples will represent the population more accurately, and their means will be closer to each other. To measure the accuracy of the sample mean we use the **standard error of the mean ($s_{\overline{X}}$)**, calculated as

$$S_{\overline{X}} = \frac{s}{\sqrt{n}}$$

where *s* is the standard deviation and $\sqrt{n}$ is the square root of the sample size. Because the standard error of the mean is computed by dividing *s* by $\sqrt{n}$, it is always a smaller value than the standard deviation.

TABLE 25.2 Sample Inclusion for Various *s* Values

Multiples of *s*	Sample Included (%)
$\overline{X} \pm 1s$	68.3
$\overline{X} \pm 1.96s$	95.0
$\overline{X} \pm 2s$	95.5
$\overline{X} \pm 3s$	99.7

Covariance and Correlation Coefficient

Often geneticists working with quantitative traits find they have to consider two phenotypic characters simultaneously. For example, a poultry breeder might investigate the correlation between body weight and egg production in hens: Do heavier birds tend to lay more eggs? The **covariance** statistic measures how much variation is common to both quantitative traits. It is calculated by taking the deviations from the mean for each trait (just as we did for estimating variance) for each individual in the sample. This gives a pair of values for each individual. The two values are multiplied together, and the sum of all these individual products is then divided by one fewer than the number in the sample. Thus, the covariance **cov_{XY}** of two sets of trait measurements, *X* and *Y*, is calculated as

$$\mathrm{cov}_{XY} = \frac{\Sigma[(X_i - \overline{X})(Y_i - \overline{Y})]}{n - 1}$$

The covariance can then be standardized as yet another statistic, the **correlation coefficient (*r*)**. The calculation of *r* is

$$r = \mathrm{cov}_{XY}/s_X s_Y$$

where s_X is the standard deviation of the first set of quantitative measurements *X*, and s_Y is the standard deviation of the second set of quantitative measurements *Y*. Values for the correlation coefficient *r* can range from -1 to $+1$. Positive *r* values mean that an increase in measurement for one trait tends to be associated with an increase in measurement for the other, while negative *r* values mean that increases in one trait are associated with decreases in the other. Therefore, if heavier hens do tend to lay more eggs, a positive *r* value can be expected. A negative *r* value, on the other hand, suggests that greater egg production is more likely from less heavy birds. One important point to note about correlation coefficients is that even significant *r* values—close to $+1$ or -1—do not prove that a cause-and-effect relationship exists between two traits. Correlation analysis simply tells us the extent to which variation in one quantitative trait is associated with variation in another, not what causes that variation.

Analysis of a Quantitative Character

To apply these statistical concepts, let's consider a genetic experiment that crossed two different homozygous varieties of tomato. One of the tomato varieties produces fruit averaging 18 oz in weight, whereas fruit from the other averages 6 oz. The F_1 obtained by crossing these two varieties has fruit weights ranging from 10 to 14 oz. The F_2 population contains individuals that produce fruit ranging from 6 to 18 oz. The results characterizing both generations are shown in **Table 25.3**.

TABLE 25.3 Distribution of F_1 and F_2 Progeny Derived from a Theoretical Cross Involving Tomatoes

	Weight (oz)												
	6	7	8	9	10	11	12	13	14	15	16	17	18
F_1 Progeny					4	14	16	12	6				
F_2 Progeny	1	1	2	0	9	13	17	14	7	4	3	0	1

The mean value for the fruit weight in the F_1 generation can be calculated as

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{626}{52} = 12.04$$

The mean value for fruit weight in the F_2 generation is calculated as

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{872}{72} = 12.11$$

Although these mean values are similar, the frequency distributions in Table 25.3 show more variation in the F_2 generation. The range of variation can be quantified as the sample variance s^2, calculated, as we saw earlier, as the sum of the squared differences between each value and the mean, divided by one less than the total number of observations.

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1}$$

When the above calculation is made, the variance is found to be 1.29 for the F_1 generation and 4.27 for the F_2 generation. When converted to the standard deviation $(s = \sqrt{s^2})$, the values become 1.13 and 2.06, respectively. Therefore, the distribution of tomato weight in the F_1 generation can be described as 12.04 ± 1.13, and in the F_2 generation it can be described as 12.11 ± 2.06.

Assuming that both tomato varieties are homozygous at the loci of interest and that the alleles controlling fruit weight act additively, we can estimate the number of polygenes involved in this trait. Since 1/72 of the F_2 offspring have a phenotype that overlaps one of the parental strains (72 total F_2 offspring; one weighs 6 oz, one weighs 18 oz; see Table 25.3), the use of the formula $1/4^n = 1/72$ indicates that n is between 3 and 4, providing evidence of the number of genes that control fruit weight in these tomato strains.

NOW SOLVE THIS

25.2 The following table shows measurements for fiber lengths and fleece weight in a small flock of eight sheep.

	Sheep Fiber Length (cm)	Fleece Weight (kg)
1	9.7	7.9
2	5.6	4.5
3	10.7	8.3
4	6.8	5.4
5	11.0	9.1
6	4.5	4.9
7	7.4	6.0
8	5.9	5.1

(*Continued*)

(a) What are the mean, variance, and standard deviation for each trait in this flock?

(b) What is the covariance of the two traits?

(c) What is the correlation coefficient for fiber length and fleece weight?

(d) Do you think greater fleece weight is correlated with an increase in fiber length? Why or why not?

■ *This problem provides data for two quantitative traits in a flock of sheep. After making numerous statistical calculations, you are asked in part (d) to determine if the traits are correlated.*

■ **HINT:** *This problem provides data for two quantitative traits and asks you to make numerous statistical calculations, ultimately determining if the traits are correlated. The key to its solution is that once the calculation of the correlation coefficient (r) is completed, you must interpret that value—whether it is positive or negative, and how close to zero it is.*

25.4 Heritability Values Estimate the Genetic Contribution to Phenotypic Variability

The question most often asked by geneticists working with multifactorial traits and diseases is how much of the observed phenotypic variation in a population is due to genotypic differences among individuals and how much is due to environment. The term **heritability** is used to describe *what proportion of total phenotypic variation in a population is due to genetic factors*. For a multifactorial trait in a given population, a high heritability estimate indicates that much of the variation can be attributed to genetic factors, with the environment having less impact on expression of the trait. With a low heritability estimate, environmental factors are likely to have a greater impact on phenotypic variation within the population.

The concept of heritability is frequently misunderstood and misused. It should be emphasized that heritability indicates neither how much of a trait is genetically determined nor the extent to which an individual's phenotype is due to genotype. In recent years, such misinterpretations of heritability for human quantitative traits have led to controversy, notably in relation to measurements such as intelligence quotients (IQs). Variation in heritability estimates for IQ among different racial groups tested led to incorrect suggestions that unalterable genetic factors control differences in intelligence levels among humans of different ancestries. Such suggestions misrepresented the meaning of heritability and ignored the contribution of genotype-by-environment interaction variance to phenotypic variation in a population. Moreover, heritability is not fixed for a trait. For example, a heritability estimate for egg production in a flock of chickens kept in individual cages might be high, indicating that differences in egg output among individual birds are largely due to genetic differences, as they all have very similar environments. For a different flock kept outdoors, heritability for egg production might be much lower, as variation among different birds may also reflect differences in their individual environments. Such differences could include how much food each bird manages to find and whether it competes successfully for a good roosting spot at night. Thus, a heritability estimate establishes the proportion of phenotypic variation that can be attributed to genetic variation *within a certain population in a particular environment*. If we measure heritability for the same trait among different populations in a range of environments, we frequently find that the calculated heritability values have large standard errors. This is an important point to remember when considering heritability estimates for traits in human populations. A mean heritability estimate of 0.65 for human height does not mean that your height is 65 percent due to your genes, but rather that in the populations sampled, on average, *65 percent of the overall variation in height could be explained by genotypic differences among individuals*.

With this subtle but important distinction in mind, we will now consider how geneticists divide the phenotypic variation observed in a population into genetic and environmental components. As we saw in Section 25.3, variation can be quantified as a sample variance: taking measurements of the trait in question from a representative sample of the population and determining the extent of the spread of those measurements around the sample mean. This gives us an estimate of the total **phenotypic variance (V_P)** in the population. Heritability estimates are obtained by using different experimental and statistical techniques to partition V_P into **genotypic variance (V_G)** and **environmental variance (V_E)** components.

An important factor contributing to overall levels of phenotypic variation is the extent to which individual genotypes affect the phenotype differently depending on the environment. For example, wheat variety A may yield an average of 20 bushels an acre on poor soil, while variety B yields an average of 17 bushels. On good soil, variety A yields 22 bushels, while variety B averages 25 bushels an acre. There are differences in yield between the two genotypically distinct varieties, so variation in wheat yield has a genetic component. Both varieties yield more on good soil, so yield is also affected by environment. However, we also see that the two varieties do not respond to better soil conditions equally: The genotype of wheat variety B achieves a greater increase in yield on good soil than does variety A. Thus, we have differences in the interaction of genotype, with environment contributing to variation for yield in populations of wheat plants. This third component of phenotypic variation is **genotype-by-environment interaction variance ($V_{G\times E}$)** (**Figure 25.6**).

We can now summarize all the components of total phenotypic variance V_P using the following equation:

$$V_P = V_G + V_E + V_{G\times E}$$

In other words, total phenotypic variance can be subdivided into genotypic variance, environmental variance, and

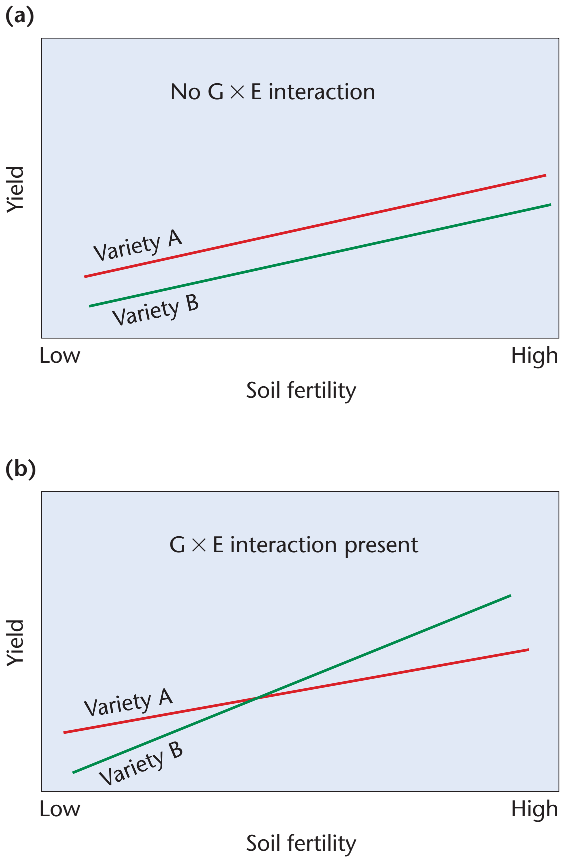


FIGURE 25.6 Differences in yield between two wheat varieties at different soil fertility levels. (a) No genotype-by-environment, or G × E, interaction: The varieties show genetic differences in yield but respond equally to increasing soil fertility. (b) G × E interaction present: Variety A outyields B at low soil fertility, but B yields more than A at high fertility levels.

genotype-by-environment interaction variance. When obtaining heritability estimates for a multifactorial trait, researchers often assume that the genotype-by-environment interaction variance is small enough that it can be ignored or combined with the environmental variance. However, it is worth remembering that this kind of approximation is another reason heritability values are *estimates* for a given population in a particular environment, not a *fixed attribute* for a trait.

Animal and plant breeders use a range of experimental techniques to estimate heritabilities by partitioning measurements of phenotypic variance into genotypic and environmental components. One approach uses inbred strains containing genetically homogeneous individuals with highly homozygous genotypes. Experiments are then designed to test the effects of a range of environmental conditions on phenotypic variability. Variation *between* different inbred strains reared in a constant environment is due predominantly to genetic factors. Variation *among* members of the same inbred strain reared under different conditions is more likely to be due to environmental factors. Other approaches involve analysis of variance for a quantitative trait among offspring from different crosses, or comparing expression of a trait among offspring and parents reared in the same environment.

Broad-Sense Heritability

Broad-sense heritability (H^2) measures the contribution of the genotypic variance to the total phenotypic variance. It is estimated as a proportion:

$$H^2 = \frac{V_G}{V_P}$$

Heritability values for a trait in a population range from 0.0 to 1.0. A value approaching 1.0 indicates that the environmental conditions have little impact on phenotypic variance, which is therefore largely due to genotypic differences among individuals in the population. Low values close to 0.0 indicate that environmental factors, not genotypic differences, are largely responsible for the observed phenotypic variation within the population studied. Few quantitative traits have very high or very low heritability estimates, suggesting that both genetics and environment play a part in the expression of most phenotypes for the trait.

The genotypic variance component V_G used in broad-sense heritability estimates includes all types of genetic variation in the population. It does not distinguish between quantitative trait loci with alleles acting additively as opposed to those with epistatic or dominance effects. Broad-sense heritability estimates also assume that the genotype-by-environment interaction variance component is negligible. While broad-sense heritability estimates for a trait are of general genetic interest, these limitations mean this kind of heritability is not very useful in breeding programs. Animal or plant breeders wishing to develop improved strains of livestock or higher-yielding crop varieties need more precise heritability estimates for the traits they wish to manipulate in a population. Therefore, another type of estimate, narrow-sense heritability, has been devised that is of more practical use.

Narrow-Sense Heritability

Narrow-sense heritability (h^2) is the proportion of phenotypic variance due to additive genotypic variance alone. Genotypic variance can be divided into subcomponents representing the different modes of action of alleles at quantitative trait loci. As not all the genes involved in a quantitative trait affect the phenotype in the same way, this partitioning distinguishes among three different kinds of gene action contributing to genotypic variance. **Additive variance (V_A)** is the genotypic variance due to the additive action of alleles at quantitative trait loci. **Dominance variance (V_D)** is the deviation from the additive components that results when phenotypic expression in heterozygotes is not precisely intermediate between the two homozygotes. **Interactive variance (V_I)** is the deviation from the additive components that occurs when two or more loci behave epistatically. The amount of interactive variance is often negligible, and so this component is often excluded from calculations of total genotypic variance.

The partitioning of the total genotypic variance V_G is summarized in the equation

$$V_G = V_A + V_D + V_I$$

and a narrow-sense heritability estimate based only on that portion of the genotypic variance due to additive gene action becomes

$$h^2 = \frac{V_A}{V_P}$$

Omitting V_I and separating V_P into genotypic and environmental variance components, we obtain

$$h^2 = \frac{V_A}{V_E + V_A + V_D}$$

Heritability estimates are used in animal and plant breeding to indicate the potential response of a population to artificial selection for a quantitative trait. Narrow-sense heritability h^2 provides a more accurate prediction of selection response than broad-sense heritability H^2 and therefore h^2 is more widely used by breeders.

Artificial Selection

Artificial selection is the process of choosing specific individuals with preferred phenotypes from an initially heterogeneous population for future breeding purposes. Theoretically, if artificial selection based on the same trait preferences is repeated over multiple generations, a population can be developed containing a high frequency of individuals with the desired characteristics. If selection is for a simple trait controlled by just one or two genes subject to little environmental influence, generating the desired population of plants or animals is relatively fast and easy. However, many traits of economic importance in crops and livestock, such as grain yield in plants, weight gain or milk yield in cattle, and speed or stamina in horses, are polygenic and frequently multifactorial. Artificial selection for such traits is slower and more complex. Narrow-sense heritability estimates are valuable to the plant or animal breeder because, as we have just seen, they estimate the proportion of total phenotypic variance for the trait that is due to additive genetic variance. Quantitative trait alleles with additive impact are those most easily manipulated by the breeder. Alleles at quantitative trait loci that generate dominance effects or interact epistatically (and therefore contribute to V_D or V_I) are less responsive to artificial selection. Thus narrow-sense heritability h^2 can be used to predict the impact of selection. The higher the estimated value for h^2 in a population, the more likely the breeder will observe a change in phenotypic range for the trait in the next generation after artificial selection.

Partitioning the genetic variance components to calculate h^2 and predict response to selection is a complex task requiring careful experimental design and analysis. The simplest approach is to select individuals with superior phenotypes for the desired quantitative trait from a heterogeneous population and breed offspring from those individuals. The mean score for the trait of those offspring ($M2$) can then be compared to that of (1) the original population's mean score (M) and (2) the selected individuals used as parents ($M1$). The relationship between these means and h^2 is

$$h^2 = \frac{M2 - M}{M1 - M}$$

This equation can be further simplified by defining $M2 - M$ as the **selection response (*R*)**, which is the degree of response to mating the selected parents, and defining $M1 - M$ as the **selection differential (*S*)**, which is the difference between the mean for the whole population and the mean for the selected population. Thus h^2 reflects the ratio of the response observed to the total response possible:

$$h^2 = \frac{R}{S}$$

A narrow-sense heritability value obtained in this way by selective breeding and measuring the response in the offspring is referred to as an estimate of **realized heritability**.

As an example of a realized heritability estimate, suppose that we measure the diameter of corn kernels in a population where the mean diameter M is 20 mm. From this population, we select a group with the smallest diameters, for which the mean $M1$ equals 10 mm. The selected plants are interbred, and the mean diameter $M2$ of the progeny kernels is 13 mm. We can calculate the realized heritability h^2 to estimate the potential for artificial selection on kernel size:

$$h^2 = \frac{M2 - M}{M1 - M}$$

$$h^2 = \frac{13 - 20}{10 - 20}$$

$$= \frac{-7}{-10}$$

$$= 0.70$$

This value for narrow-sense heritability indicates that the selection potential for kernel size is relatively high.

The longest running artificial selection experiment known is still being conducted at the State Agricultural Laboratory in Illinois. Since 1896, corn has been selected for both high and low oil content. After the initial 76 generations, selection continued to result in increased oil content (**Figure 25.7**). With each cycle of successful selection, more of the corn plants accumulate a higher percentage of additive alleles involved in oil production. Consequently, the narrow-sense heritability h^2 of increased oil content in succeeding generations has declined (see parenthetical values at generations 9, 25, 52, and 76 in Figure 25.7) as artificial selection comes closer and closer to optimizing the genetic potential for oil production. Theoretically, the process will continue

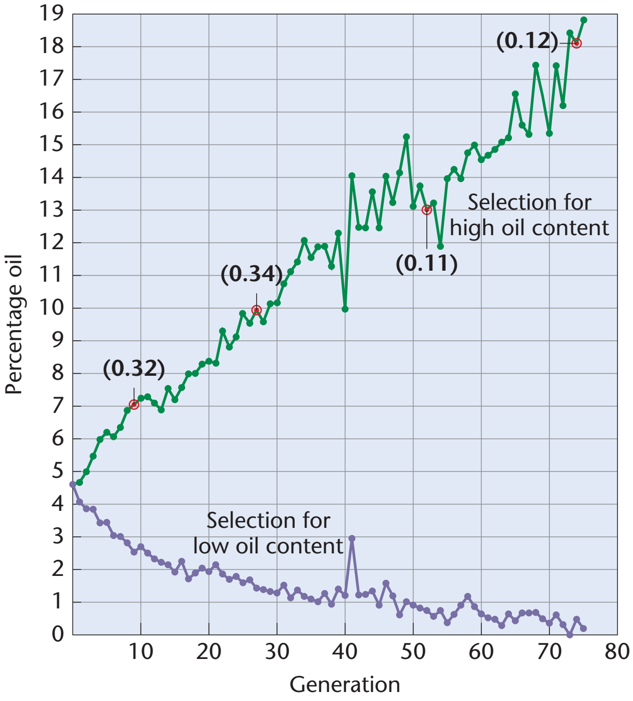


FIGURE 25.7 Response of corn selected for high and low oil content over 76 generations. The numbers in parentheses at generations 9, 25, 52, and 76 for the high-oil line indicate the calculation of heritability at these points in the continuing experiment.

until all individuals in the population possess a uniform genotype that includes all the additive alleles responsible for high oil content. At that point, h^2 will be reduced to zero and response to artificial selection will cease. The decrease in response to selection for low oil content shows that heritability for low oil content is approaching this point.

Table 25.4 lists narrow-sense heritability estimates expressed as percentage values for a variety of quantitative traits in different organisms. As you can see, these h^2 values vary, but heritability tends to be low for quantitative traits that are essential to an organism's survival. Remember, this does not indicate the absence of a genetic contribution to the observed phenotypes for such traits. Instead, the low h^2 values show that natural selection has already largely optimized the genetic component of these traits during evolution. Egg production, litter size, and conception rate are examples of how such physiological limitations on selection have already been reached. Traits that are less critical to survival, such as body weight, tail length, and wing length, have higher heritabilities because more genotypic variation for such traits is still present in the population. Remember also that any single heritability estimate can only provide information about one population in a specific environment. Therefore, narrow-sense heritability is a more valuable predictor of response to selection when estimates are calculated for many populations and environments and show the presence of a clear trend.

TABLE 25.4 Estimates of Heritability for Traits in Different Organisms

Trait	Heritability (h^2)
Mice	
Tail length	60%
Body weight	37
Litter size	15
Chickens	
Body weight	50
Egg production	20
Egg hatchability	15
Cattle	
Birth weight	45
Milk yield	44
Conception rate	3

Limitations of Heritability Studies

While the above discussion makes clear that heritability studies are valuable in estimating the genetic contribution to phenotypic variance, the knowledge gained about heritability of traits must be balanced by awareness of some of the constraints inherent in such estimates:

- Heritability studies provide no information about which specific genes influence the traits being evaluated.
- Heritability is measured in populations and has only limited application to individuals.
- Measured heritability depends on the environmental variation present in the population being studied and cannot be used to evaluate differences between populations.
- Future changes in environmental factors can affect heritability.

25.5 Twin Studies Allow an Estimation of Heritability in Humans

Human twins are useful subjects for examining how much phenotypic variance for a multifactorial trait is due to the genotype as opposed to the environment. In these studies, the underlying principle has been that **monozygotic (MZ)**, or **identical, twins** are derived from a single zygote that divides mitotically and then spontaneously splits into two separate cells. Both cells give rise to a genotypically identical embryo. **Dizygotic (DZ)**, or **fraternal, twins**, on the other hand, originate from two separate fertilization events and are only as genetically similar as any two siblings, with an average of 50 percent of their alleles in common. For a given trait, therefore, phenotypic differences between pairs of MZ twins will be equivalent to the environmental variance V_E (because the genotypic variance is zero). Phenotypic differences between DZ twins, however, display both environmental variance V_E and approximately half the genotypic variance V_G. Comparing the extent of phenotypic variance for the same trait in MZ and DZ sets of twins provides an estimate of broad-sense heritability for the trait.

Twins are said to be **concordant** for a given trait if both express it or neither expresses it. If one expresses the trait and the other does not, the pair is said to be **discordant**. Comparison of concordance values of MZ versus DZ twins reared together illustrates the potential value for heritability assessment. (See the Now Solve This feature on page 611, for example.)

Before any conclusions can be drawn from twin studies, the data must be examined carefully. For example, if concordance values approach 90 to 100 percent in MZ twins, we might be inclined to interpret that as a large genetic contribution to the phenotype of the trait. In some cases—for example, blood types and eye color—we know that this is indeed true. In the case of contracting measles, however, a high concordance value merely indicates that the trait is almost always induced by a factor in the environment—in this case, a virus.

It is more meaningful to compare the *difference* between the concordance values of MZ and DZ twins. If concordance values are significantly higher in MZ twins, we suspect a strong genetic component in the determination of the trait. In the case of measles, where concordance is high in both types of twins, the environment is assumed to be the major contributing factor. Such an analysis is useful because phenotypic characteristics that remain similar in different environments are likely to have a strong genetic component.

Large-Scale Analysis of Twin Studies

For decades, researchers have used twin studies to examine the relative contributions of genotype and environment to the phenotypic variation observed in complex traits in humans. These traits involve the interplay of multiple genes with a network of environmental factors, and the genetic components of the resulting phenotypic variance can be difficult to study. The simplest way to assess the genetic contribution is to assume that the effect of each gene on a trait is independent of the effects of other genes. Because the effects of all genes are added together, this is called the *additive model*. However, in recent years, some geneticists have proposed that nonadditive factors such as dominance and epistasis are more important than additive genetic effects. As a result, the relative roles of additive and nonadditive factors are a subject of active debate.

In an attempt to resolve this issue, an international project recently examined the results of all twin studies performed in the last 50 years. This study, published in 2015, involved the compilation and analysis of the data for over 17,000 traits studied in more than 14 million twin pairs drawn from more than 2700 published papers.

Several important general conclusions can be drawn from this landmark study. First, based on correlations between MZ and DZ twin pairs, which can be used to draw conclusions about how likely it is that genetic influences on a trait are mostly additive or nonadditive, researchers concluded that the vast majority of traits follow a simple additive model, providing strong support for one of the foundations of heritability studies. This does not exclude the role of nonadditive factors such as dominance and epistasis, but these factors most likely play a secondary role in heritability. Second, the results are consistent with the findings from genome-wide association studies (GWAS) that many complex traits are controlled by many genes, each with a small effect. Third, genetic variance is an important component of the individual variations observed in populations. In addition, the relative effects of genotypes and environmental factors are nonrandomly distributed, making their contributions somewhat trait-specific.

The data from this study are available in a Web-based application, Meta-analysis of Twin Correlations and Heritability (MaTCH), which can be used as a resource for the study of complex traits and the genetic and environmental components of heritability.

Twin Studies Have Several Limitations

Interesting as they are, human twin studies contain some unavoidable sources of error. For example, MZ twins are often treated more similarly by parents and teachers than are DZ twins, especially when the DZ siblings are of a different sex. This circumstance may inflate the environmental variance for DZ twins. Another possible error source is interactions between the genotype and the environment that produce variability in the phenotype. These interactions can increase the total phenotypic variance for DZ twins compared to MZ twins raised in the same environment, influencing heritability calculations. Overall, heritability estimates for human traits based on twin studies should therefore be considered approximations and examined very carefully before any conclusions are drawn.

Although they must often be viewed with caution, classical twin studies, based on the assumption that MZ twins share the same genome, have been valuable for estimating heritability over a wide range of traits including multifactorial disorders such as cardiovascular disease, diabetes, and mental illness, for example. These disorders clearly have genetic components, and twin studies provide a foundation for studying interactions between genes and environmental factors. However, results from genomics research have challenged the view that MZ twins are truly identical and have forced a reevaluation of both the methodology and the results of twin studies. Such research has also opened the way to new approaches to the study of interactions between the genotype and environmental factors.

The most relevant genomic discoveries about twins include the following:

- By the time they are born, MZ twins do not necessarily have identical genomes.
- Gene-expression patterns in MZ twins change with age, leading to phenotypic differences.

We will address these points in order. First, MZ twins develop from a single fertilized egg, where sometime early in development the resulting cell mass separates into two distinct populations creating two independent embryos. Until that time, MZ twins have identical genotypes. Subsequently, however, the genotypes can diverge slightly. For example, differences in *copy number variation* (*CNV*)—variation in the number of copies of numerous large DNA sequences (usually 1000 bp or more)—may arise, differentially producing genetically distinct populations of cells in each embryo (see Chapter 8 for a discussion of CNV). This creates a condition called *somatic mosaicism*, which may result in a milder disease phenotype in some disorders and may play a similar role in phenotypic discordance observed in some pairs of MZ twins.

At this point, it is difficult to know for certain how often CNV arises after MZ twinning, but one estimate suggests that such differences are believed to occur in 10 percent of all twin pairs. In those pairs where it does occur, one estimate is that such divergence takes place in 15 to 70 percent of the somatic cells. In one case, a CNV difference between MZ twins has been associated with chronic lymphocytic leukemia in one twin, but not the other.

The second genomic difference between MZ twins involves **epigenetics**—the chemical modification of their DNA and associated histones. An international study of epigenetic modifications in adult European MZ twins showed that MZ twin pairs are epigenetically identical at birth, but adult MZ twins show significant differences in the *methylation patterns* of both DNA and histones. Such epigenetic changes in turn affect patterns of gene expression. The accumulation of epigenetic changes and gene-expression profiles may explain some of the observed phenotypic discordance and susceptibility to diseases in adult MZ twins. For example, a clear difference in DNA methylation patterns is observed in MZ twins discordant for Beckwith–Wiedemann syndrome, a genetic disorder associated with variable developmental overgrowth of certain tissues and organs and an increased risk of developing cancerous and noncancerous tumors. Affected infants are often larger than normal, and one in five die early in life.

Other complex disorders displaying a genetic component are similarly being investigated using epigenetic analysis in twin studies. These include susceptibility to several neurobiological disorders, including schizophrenia and autism, as well as to the development of Type I diabetes, breast cancer, and autoimmune disease.

Progressive, age-related genomic modifications may be the result of MZ twins being exposed to different environmental factors, or from failure of epigenetic marking following DNA replication. These findings also indicate that concordance studies in DZ twins must take into account genetic as well as *epigenetic differences* that contribute to discordance in these twin pairs.

The realization that epigenetics may play an important role in the development of phenotypes promises to make twin studies an especially valuable tool in dissecting the interactions among genes and the role of environmental factors in the production of phenotypes. Once the degree of epigenetic differences between MZ and DZ twin pairs has been defined, molecular studies on DNA and histone modification can link changes in gene expression with differences in the concordance rates between MZ and DZ twins.

We will discuss the most recent findings involving epigenetics and summarize its many forms and functions later in the text (see Chapter 19—Epigenetic Regulation in Eukaryotes).

NOW SOLVE THIS

25.3 The following table gives the percentage of twin pairs studied in which both twins expressed the same phenotype for a trait (concordance). Percentages listed are for concordance for each trait in monozygotic (MZ) and dizygotic (DZ) twins. Assuming that both twins in each pair were raised together in the same environment, what do you conclude about the relative importance of genetic versus environmental factors for each trait?

Trait	MZ %	DZ %
Blood types	100	66
Eye color	99	28
Mental retardation	97	37
Measles	95	87
Hair color	89	22
Handedness	79	77
Idiopathic epilepsy	72	15
Schizophrenia	69	10
Diabetes	65	18
Identical allergy	59	5
Cleft lip	42	5
Club foot	32	3
Mammary cancer	6	3

■ **HINT:** *This problem asks you to evaluate the relative importance of genetic versus environmental contributions to specific traits by examining concordance values in MZ versus DZ twins. The key to its solution is to examine the difference in concordance values and to factor in what you have learned about the genetic differences between MZ and DZ twins.*

25.6 Quantitative Trait Loci Are Useful in Studying Multifactorial Phenotypes

Environmental effects, interaction among segregating alleles, and the large number of genes that may contribute to a polygenic phenotype make it difficult to (1) identify all genes that are involved and (2) determine the effect of each gene on the phenotype. However, because many quantitative traits are of economic or medical relevance, it is often desirable to obtain this information. In such studies, a chromosome region is identified as containing one or more genes contributing to a quantitative trait known as a

quantitative trait locus (**QTL**, or **QTLs** if plural). When possible, the relevant gene or genes contained within a QTL are isolated and studied.

The modern approach used to find and map QTLs involves looking for associations between DNA markers and phenotypes. One way to do this is to begin with individuals from two lines created by artificial selection that are highly divergent for a phenotype (fruit weight, oil content, bristle number, etc.). For example, **Figure 25.8** illustrates a generic case of QTL mapping. Over many generations of artificial selection, two divergent lines become highly homozygous, which facilitates their use in QTL mapping. Individuals from each of the lines with divergent phenotypes [generation 25 in **Figure 25.8(a)**] are used as parents to create an F_1 generation whose members will be heterozygous at most of the loci contributing to the trait. Additional crosses, either among F_1 individuals or between the F_1 and the inbred parent lines, result in F_2 generations that carry different portions of the parental genomes [**Figure 25.8(b)**] with different QTL genotypes and associated phenotypes. This segregating F_2 is known as the **QTL mapping population**.

Researchers then measure phenotypic expression of the trait among individuals in the mapping population and identify genomic differences among individuals by using chromosome-specific DNA markers such as *restriction fragment length polymorphisms* (*RFLPs*), *microsatellites*, and *single-nucleotide polymorphisms* (*SNPs*) (see Chapter 21). Computer-based statistical analysis is used to search for linkage between the markers and a component of phenotypic variation associated with the trait. If a DNA marker (such as those markers described previously) *is not* linked to a QTL, then the phenotypic mean score for the trait will not vary among individuals with different genotypes at that marker locus. However, if a DNA marker *is* linked to a QTL, then different genotypes at that marker locus will also differ in their phenotypic expression of the trait. When this occurs, the marker locus and the QTL are said to *cosegregate*. Consistent cosegregation establishes the presence of a QTL at or near the DNA marker along the chromosome—in other words, the marker and QTL are linked. When numerous QTLs for a given trait have been located, a genetic map

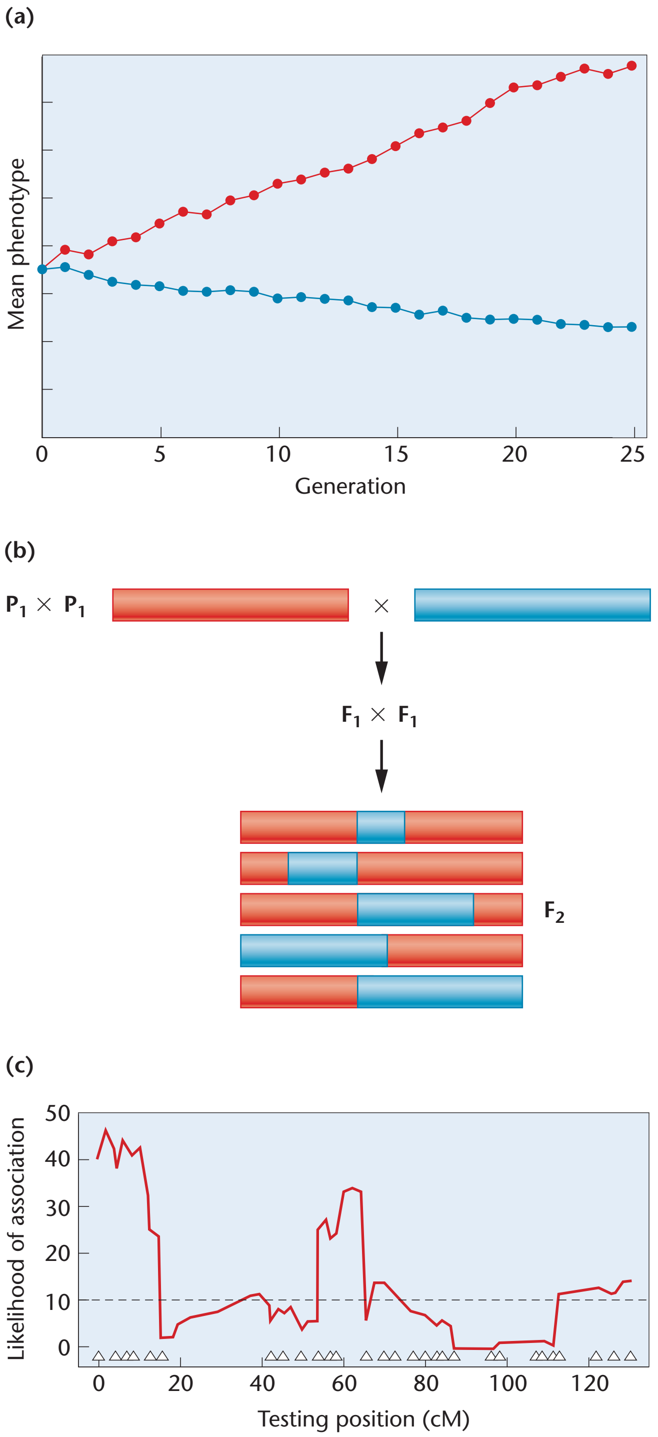


FIGURE 25.8 (a) Individuals from highly divergent lines created by artificial selection are chosen from generation 25 as parents. (b) The thick bars represent the genomes of individuals selected from the divergent lines as parents. These individuals are crossed to produce an F_1 generation (not shown). An F_2 generation is produced by crossing members of the F_1. As a result of crossing over, individual members of the F_2 generation carry different portions of the P_1 genome, as shown by the colored segments of the thick bars. DNA markers and phenotypes in individuals of the F_2 generation are analyzed. (c) Statistical methods are used to determine the probability that a DNA marker is associated with a QTL that affects the phenotype. The results are plotted as the likelihood of association against chromosomal location. Units on genetic maps are measured in centimorgans (cM), determined by crossover frequencies. Peaks above the horizontal line represent significant results. The data show five possible QTLs, with the most significant findings at about 10 cM and 60 cM.

is created, showing the probability that specific chromosomal regions are associated with the phenotype of interest [**Figure 25.8(c)**]. Further research using genomic techniques identifies genes in these regions that contribute to the phenotype.

QTL mapping has been used extensively in agriculture, including for plants such as corn, rice, wheat, and tomatoes (**Table 25.5**), and livestock such as cattle, pigs, sheep, and chickens.

Tomatoes are one of the world's major agricultural crops, and several hundred varieties are grown and harvested each year. To aid in the creation of new varieties, hundreds of QTLs for traits including fruit size, shape, soluble solid content, and acidity have been identified and mapped to all 12 chromosomes in the tomato genome. In addition, the genomes of several tomato varieties have been sequenced. We will describe studies focused on quantitative traits controlling fruit shape and weight as an example of QTL research.

While the cultivated tomato can weigh up to 1000 grams, fruit from the related wild species thought to be the ancestor of the modern tomato weighs only a few grams (see **Figure 25.9**). In a study by Steven Tanksley and colleagues, QTL mapping has identified more than 28 QTLs related to this thousand-fold variation in fruit weight. More than ten years of work was required to localize, identify, and clone one of these QTLs, called *fw2.2* (on chromosome 2). Within this QTL, a specific gene, *ORFX*, has been identified, and alleles at this locus are responsible for about 30 percent of the variation in fruit weight.

The *ORFX* gene has been isolated, cloned, and transferred between plants, with interesting results. One allele of *ORFX* is present in all wild small-fruited varieties of tomatoes investigated, while another allele is present in all domesticated large-fruited varieties. When a cloned *ORFX* gene from small-fruited varieties is transferred to a plant that normally produces large tomatoes, the transformed plant produces fruits that are greatly reduced in weight. In the varieties studied by Tanksley's group, the reduction averaged 17 grams, a statistically significant phenotypic change caused by the action of a gene found within a single QTL.

TABLE 25.5 QTLs for Quantitative Phenotypes

Organism	Quantitative Phenotype	QTLs Identified
Tomato	Soluble solids	7
	Fruit mass	13
	Fruit pH	9
	Growth	5
	Leaflet shape	9
	Height	9
Maize	Height	11
	Leaf length	7
	Grain yield	18
	Number of ears	10

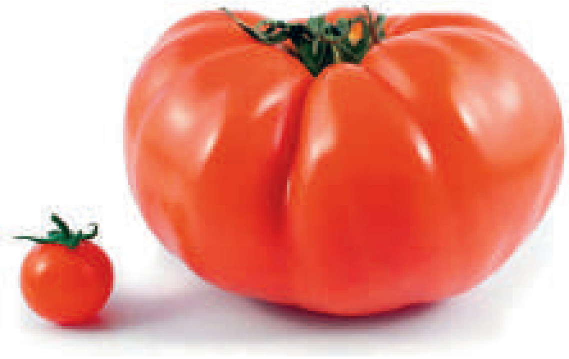


FIGURE 25.9 A wild species of tomato, similar in size to the tomato on the left, is regarded as the ancestor of all modern tomatoes, including the beefsteak tomato shown at the right.

Further analysis of *ORFX* revealed that this gene encodes a protein that negatively regulates cell division during fruit development. Differences in the time of gene expression and differences in the amount of transcript produced lead to small or large fruit. Higher levels of expression mediated by transferred *ORFX* alleles exert a negative control over cell division, resulting in smaller tomatoes.

Yet *ORFX* and other related genes cannot account for all the observed variation in tomato size. Analysis of two other QTLs, *lc* (located on chromosome 2) and *fas* (on chromosome 11), indicates that the development of extreme differences in fruit size resulting from artificial selection also involves an increase in the number of seed compartments, called locules, in the mature fruit. Small, ancestral varieties produce fruit with two to four locules, but the large-fruited present-day strains have six or more of these compartments. The *lc* QTL maps to a noncoding region of the genome and consists of two SNPs that regulate the expression of nearby genes responsible for some of the increase in locule number. In addition, *lc* interacts with certain alleles of *fas*, another SNP locus, to further increase locule number, giving rise to the wide range of sizes and shapes in present-day tomatoes (**Figure 25.10**). Thus, QTLs that affect fruit size in tomatoes work by controlling at least two developmental processes: cell division early in development (*ORFX*) and the determination of the number of seed compartments (*lc* and *fas*).

Expression QTLs Regulate Gene Expression

The discovery that QTLs can control gene expression led researchers to systematically hunt for genomic loci that regulate the expression of one or more genes involved in

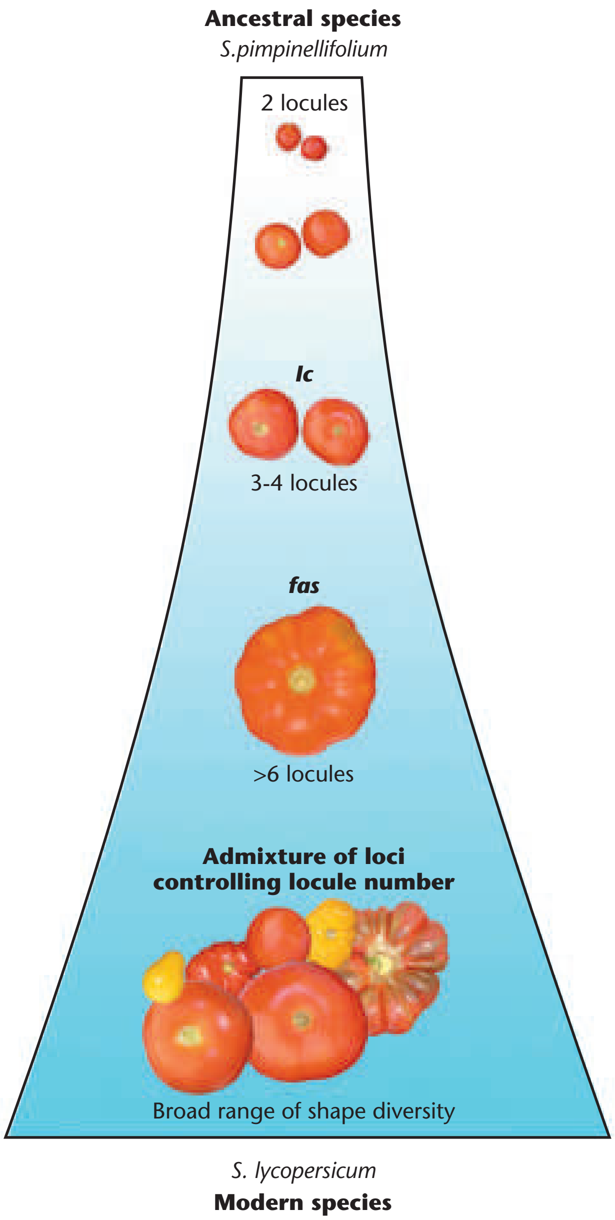


FIGURE 25.10 Changes in locule number during tomato domestication. The ancestral species *S. pimpinellifolium* contains two locules. At some point, a high-locule allele of *lc* was introduced and probably appeared before the introduction of the present-day *fas* allele, which further expanded locule number. These two QTLs are the major loci controlling locule number. As alleles of other loci controlling locule number were introduced into domesticated varieties, phenotypic diversity in the modern-day species *S. lycopersicum* expanded even further.

quantitative traits. These loci, called **expression QTLs (eQTLs)**, can be identified with the same methods used to find other QTLs. The first eQTLs were discovered in yeast but have now been identified in the genomes of many plants and animals. However, eQTLs differ from more classical QTLs in that the phenotype of variable gene expression can be regulated at any of the many steps along the path from DNA to protein (see Chapters 17 and 18 for a discussion of gene regulation in eukaryotes).

For example, most eQTLs identified to date are noncoding genomic variants, including SNPs or short indels (insertions/deletions) that affect transcription factor (TF) binding, the action of promoters and enhancers, and pre-mRNA processing. DNA sequence variations located in TF-binding sites are not only associated with differences in binding efficiency but are also linked to changes in DNA methylation, mRNA levels, and nucleosomal and chromatin changes including histone modifications. Figure 25.11 summarizes the molecular regulatory pathways used by eQTLs that result in variable gene expression and phenotypic variation.

Expression QTLs and Genetic Disorders

We conclude this chapter by discussing the role that variation in the level of gene expression plays in the phenotypic variation observed in complex disorders. In humans, variation resulting from the action of eQTLs encompasses a wide spectrum of phenotypes ranging from normal variations to disease states. The ability to study the expression of eQTLs and gene variability in the same individual helped identify the association between genes and disease and the network of genes controlling those disorders. This approach has identified genes responsible for complex diseases such as asthma, cleft lip, Type 2 diabetes, and coronary artery disease. Asthma cases have risen dramatically over the last three decades, and this disease is now a major public health concern. Genome-wide association studies (GWAS) have identified loci that confer susceptibility to asthma; however, the functions of many of these genes are unknown, and GWAS alone are unable to establish which alleles of these loci are responsible for susceptibility or the mechanism of their action.

To identify genes directly involved in asthma susceptibility, researchers collected lung specimens from over 1000 individuals and used lung-specific gene expression as a phenotype to study how genetic variants (DNA polymorphisms) are linked to both gene expression (eQTLs) and the asthma phenotype. Integration of GWAS and eQTL data identified 34 genes that form six interconnected networks that constitute the gene set that causes asthma. Each network contains a single driver gene that controls the other genes in each network. These six driver genes are now candidates for drug discovery studies to develop therapies for this chronic and sometimes fatal disease. Similar approaches are likely to reveal the genetic networks that underlie other complex genetic disorders.

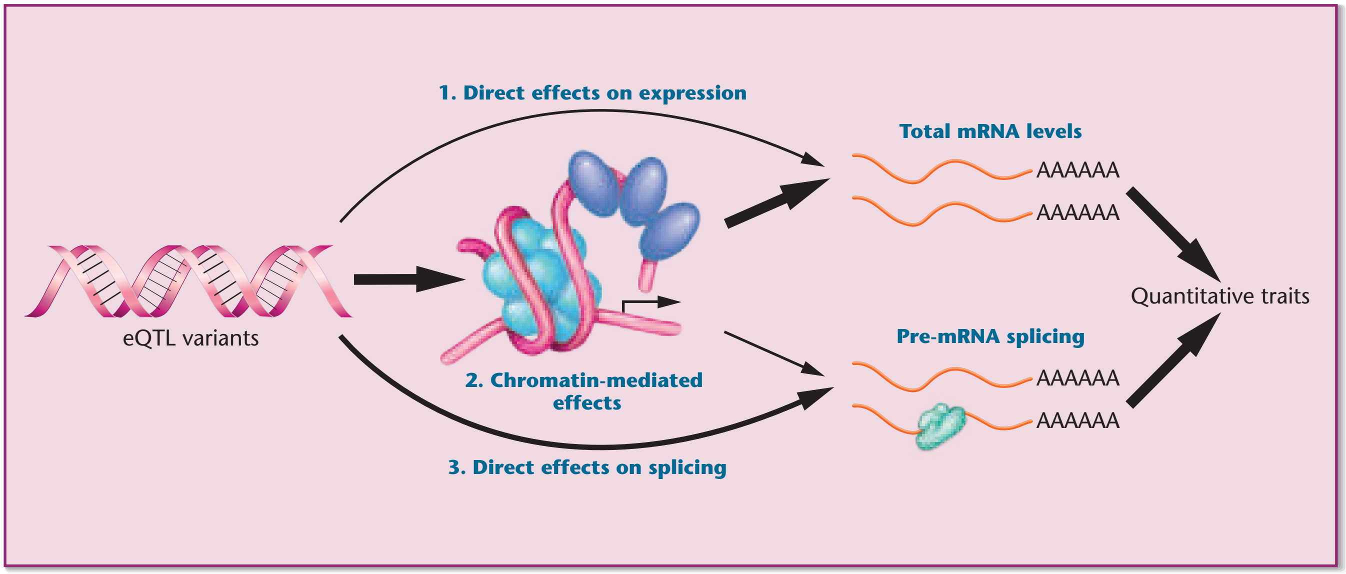


FIGURE 25.11 eQTL variants control gene expression through three major pathways: (1) direct effects on expression that control the amount of mRNA produced, (2) chromatin-mediated effects, including transcription factor binding, histone modification, and DNA methylation, and (3) direct effects on splicing of pre-mRNA, producing variant mRNA molecules. The contribution of each pathway is indicated by the thickness of arrows. Over 60 percent of eQTL variants regulate gene expression via chromatin modifications.

GENETICS, ETHICS, AND SOCIETY

Rice, Genes, and the Second Green Revolution

Of the 7 billion people now living on Earth, more than 800 million do not have enough to eat. This number is expected to grow by an additional 1 million people each year for the next several decades. How will we be able to feed the estimated 8 billion people on Earth by 2025?

The past gives us some basis for optimism. In the 1950s and 1960s, plant scientists set about to increase the production of crop plants, including the three most important grains—rice, wheat, and maize. These efforts became known as the *Green Revolution*. The approach was three-fold: (1) to increase the use of fertilizers, pesticides, and irrigation; (2) to bring more land under cultivation; and (3) to develop improved varieties of crop plants by intensive plant breeding. The results were dramatic. Developing nations more than doubled their production of rice, wheat, and maize between 1961 and 1985. The Green Revolution saved millions of people from starvation and improved the quality of life for millions more; however, its impact is beginning to wane. If food production is to keep pace with the projected increase in the world's population, we will need a second Green Revolution.

Central to the success of the Green Revolution was research involving rice, upon which about half of the Earth's population depends. Advances were facilitated by the establishment in 1960 of the International Rice Research Institute (IRRI). One of their major accomplishments was the creation of a rice variety with improved disease resistance and higher yields. The IRRI research team crossed a Chinese rice variety (*Dee-geo-woo-gen*) and an Indonesian variety (*Peta*) to create a new cultivar known as IR8. IR8 produced a greater number of rice kernels per plant. However, IR8 plants were so top heavy with grain that they tended to fall over—a trait called "lodging." To reduce lodging, IRRI breeders crossed IR8 with a dwarf native variety to create semi-dwarf lines. Due in part to the adoption of the semi-dwarf IR8 lines, the world production of rice doubled in 25 years.

Despite the progress brought about by the introduction of IR8, scientists predict that we will need a further 40 percent increase in the annual rice harvest to keep pace with anticipated population growth during the next 30 years. Not only will we need higher yields, but also new rice varieties with greater disease resistance and tolerance

(continued)

Genetics, Ethics, and Society, continued

to extreme climate changes, drought, salinity, and loss of soil fertility. Dozens of quantitative trait loci (QTLs) appear to contribute to these traits, making the task even more challenging. In the near future, scientists will need to introduce these traits into current dwarf varieties of domestic rice, using conventional breeding, genomics and genetic engineering.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references to help you discuss some of the technical and ethical issues surrounding the Green Revolution.

1. Scientists from IRRI and other research centers are working to develop new rice varieties. Describe several of these new varieties and how they may contribute to the second Green Revolution.

   *Learn about some of the research projects supported by IRRI on the IRRI Web site (*http://irri.org/our-work/our-research-networks*). Also, read about the C4 project at* Dayton, L. (2014). Blue-sky rice. *Nature* 514:S52–S54.

2. Despite its benefits, some critics question the long-term practical and ethical outcomes of the Green Revolution. What are some of these questions and criticisms? Which of these do you think has merit, and how can some of them be addressed?

   A discussion of this topic can be found in Pingali, P. L. (2012). Green Revolution: Impacts, limits, and the path ahead. *Proc. Natl. Acad. Sci. USA* 109 (31):12302–12308. *Also see* Ellison, K., and Wellner, K. (2016). The Green Revolution: Research, Ethics, and Society at http://www.onlineethics.org/Resources/Cases/GreenRevolution.aspx.

CASE STUDY A Chance Discovery

At an interview with a genetic counselor, a couple with a severely asthmatic child learned that asthma is a complex disorder involving many genetic loci. The counselor explained that a method called whole genome sequencing (WGS) is now widely used in diagnosing and treating traits controlled by multiple loci and, in this case, could provide information to devise an effective therapy for their child. However, the parents were warned that because their child's entire genome was to be sequenced, information unrelated to asthma, but with potentially serious health consequences, might be discovered. After permission was granted, genome analysis created a panel of loci for therapy design. The analysis also revealed that the child carried two copies of an allele conferring an increased risk for Alzheimer disease. One copy of this allele increases the risk 4-fold; two copies raise the risk to 12-fold. Even though the child and both parents are at risk, current guidelines do not require that this finding be disclosed because it is unrelated to the primary reason for undertaking WGS. Knowing that disclosure was not legally required, but feeling she may have an ethical responsibility to divulge this information, the counselor was conflicted regarding how to proceed.

1. Based on the outcome of the WGS, what can the counselor tell the parents about their own risk of developing Alzheimer disease?
2. If you were the counselor, would you disclose this information to the parents, and if so, what is your reasoning?
3. Would the fact that there is currently no treatment for Alzheimer disease influence your decision about disclosure?

See Roche, M., and Berg, J. (2015). Incidental findings with genomic testing: Implications for genetic counseling practice. *Curr. Genet. Med. Rep.* 3:166–176.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Quantitative inheritance results in a range of phenotypes due to the action of additive alleles from two or more genes, as influenced by environmental factors.
2. Numerous statistical methods are essential during the analysis of quantitative traits, including the mean, variance, standard deviation, standard error, covariance, and the correlation coefficient.
3. Heritability is an estimate of the relative contribution of genetic versus environmental factors to the range of phenotypic variation seen in a quantitative trait in a particular population and environment.
4. Twin studies, while having some limitations, are useful in assessing heritabilities for polygenic traits in humans.
5. Quantitative trait loci (QTLs) may be identified and mapped using DNA markers. In turn, other loci, expression QTLs (eQTLs) regulate transcription in QTLs.

INSIGHTS AND SOLUTIONS

1. In a certain plant, height varies from 6 to 36 cm. When 6-cm and 36-cm plants were crossed, all F_1 plants were 21 cm. In the F_2 generation, a continuous range of heights was observed. Most were around 21 cm, and 3 of 200 were as short as the 6-cm P_1 parent.

   (a) What mode of inheritance does this illustrate, and how many gene pairs are involved?

   (b) How much does each additive allele contribute to height?

   (c) List all genotypes that give rise to plants that are 31 cm.

   Solution:

   (a) Polygenic inheritance is illustrated when a trait is continuous and when alleles contribute additively to the phenotype. The 3/200 ratio of F_2 plants is the key to determining the number of gene pairs. This reduces to a ratio of 1/66.7, very close to 1/64. Using the formula $1/4^n = 1/64$ (where 1/64 is equal to the proportion of F_2 phenotypes as extreme as either P_1 parent), $n = 3$. Therefore, three gene pairs are involved.

   (b) The variation between the two extreme phenotypes is

$$36 - 6 = 30 \text{ cm}$$

Because there are six potential additive alleles (*AABBCC*), each contributes

$$30/6 = 5 \text{ cm}$$

to the base height of 6 cm, which results when no additive alleles (*aabbcc*) are part of the genotype.

   (c) All genotypes that include five additive alleles will be 31 cm (5 alleles × 5 cm/allele + 6 cm base height = 31 cm). Therefore, *AABBCc*, *AABbCC*, and *AaBBCC* are the genotypes that will result in plants that are 31 cm.

2. In a cross separate from the above-mentioned F_1 crosses, a plant of unknown phenotype and genotype was testcrossed, with the following results:

1/4	11 cm
2/4	16 cm
1/4	21 cm

   An astute genetics student realized that the unknown plant could be only one phenotype but could be any of three genotypes. What were they?

   Solution: When testcrossed (with *aabbcc*), the unknown plant must be able to contribute either one, two, or three additive alleles in its gametes in order to yield the three phenotypes in the offspring. Since no 6-cm offspring are observed, the unknown plant never contributes all nonadditive alleles (*abc*). Only plants that are homozygous at one locus and heterozygous at the other two loci will meet these criteria. Therefore, the unknown parent can be any of three genotypes, all of which have a phenotype of 26 cm:

AABbCc
AaBbCC
AaBBCc

   For example, in the first genotype (*AABbCc*),

$$AABbCc \times aabbcc$$

   yields

1/4	*AaBbCc*	21 cm
1/4	*AaBbcc*	16 cm
1/4	*AabbCc*	16 cm
1/4	*Aabbcc*	11 cm

   which is the ratio of phenotypes observed.

3. The mean and variance of corolla length in two highly inbred strains of *Nicotiana* and their progeny are shown in the following table. One parent (P_1) has a short corolla, and the other parent (P_2) has a long corolla. Calculate the broad-sense heritability (H^2) of corolla length in this plant.

Strain	Mean (mm)	Variance (mm)
P_1 short	40.47	3.12
P_2 long	93.75	3.87
F_1 ($P_1 \times P_2$)	63.90	4.74
F_2 ($F_1 \times F_1$)	68.72	47.70

   Solution: The formula for estimating heritability is $H^2 = V_G/V_P$, where V_G and V_P are the genetic and phenotypic components of variation, respectively. The main issue in this problem is obtaining some estimate of two components of phenotypic variation: genetic and environmental factors. V_P is the combination of genetic and environmental variance. Because the two parental strains are true breeding, they are assumed to be homozygous, and the variance of 3.12 and 3.87 is considered to be the result of environmental influences. The average of these two values is 3.50. The F_1 is also genetically homogeneous and gives us an additional estimate of the impact of environmental factors. By averaging this value along with that of the parents,

$$\frac{4.74 + 3.50}{2} = 4.12$$

   we obtain a relatively good idea of environmental impact on the phenotype. The phenotypic variance in the F_2 is the sum of the genetic (V_G) and environmental (V_E) components. We have estimated the environmental input as 4.12, so 47.70 minus 4.12 gives us an estimate of V_G of 43.58. Heritability then becomes 43.58/47.70, or 0.91. This value, when interpreted as a percentage, indicates that about 91 percent of the variation in corolla length is due to genetic influences.

Problems and Discussion Questions

Mastering Genetics Visit for instructor-assigned tutorials and problems.

1. **HOW DO WE KNOW?** In this chapter, we focused on a mode of inheritance referred to as quantitative genetics, as well as many of the statistical parameters utilized to study quantitative traits. Along the way, we found opportunities to consider the methods and reasoning by which geneticists acquired much of their understanding of quantitative genetics. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
(a) How do we know that threshold traits are actually polygenic even though they may have as few as two discrete phenotypic classes?
(b) How can we ascertain the number of polygenes involved in the inheritance of a quantitative trait?
(c) What findings led geneticists to postulate the multiple-factor hypothesis that invoked the idea of additive alleles to explain inheritance patterns?
(d) How do we assess environmental factors to determine if they impact the phenotype of a quantitatively inherited trait?
(e) How do we know that monozygotic twins are not identical genotypically as adults?

2. **CONCEPT QUESTION** Review the Chapter Concepts list on page 599. These all center around quantitative inheritance and the study and analysis of polygenic traits. Write a short essay that discusses the difference between the more traditional Mendelian and neo-Mendelian modes of inheritance (qualitative inheritance) and quantitative inheritance.

3. Define the following: (a) polygenic, (b) additive alleles, (c) correlation, (d) monozygotic and dizygotic twins, (e) heritability, (f) QTL, and (g) continuous variation.

4. A dark-red strain and a white strain of wheat are crossed and produce an intermediate, medium-red F_1. When the F_1 plants are interbred, an F_2 generation is produced in a ratio of 1 dark-red: 4 medium-dark-red: 6 medium-red: 4 light-red: 1 white. Further crosses reveal that the dark-red and white F_2 plants are true breeding.
(a) Based on the ratios in the F_2 population, how many genes are involved in the production of color?
(b) How many additive alleles are needed to produce each possible phenotype?
(c) Assign symbols to these alleles, and list possible genotypes that give rise to the medium-red and light-red phenotypes.
(d) Predict the outcome of the F_1 and F_2 generations in a cross between a true-breeding medium-red plant and a white plant.

5. Height in humans depends on the additive action of genes. Assume that this trait is controlled by the four loci *R*, *S*, *T*, and *U* and that environmental effects are negligible. Instead of additive versus nonadditive alleles, assume that additive and partially additive alleles exist. Additive alleles contribute two units, and partially additive alleles contribute one unit to height.
(a) Can two individuals of moderate height produce offspring that are much taller or shorter than either parent? If so, how?
(b) If an individual with the minimum height specified by these genes marries an individual of intermediate or moderate height, will any of their children be taller than the tall parent? Why or why not?

6. An inbred strain of plants has a mean height of 24 cm. A second strain of the same species from a different geographic region also has a mean height of 24 cm. When plants from the two strains are crossed together, the F_1 plants are the same height as the parent plants. However, the F_2 generation shows a wide range of heights; the majority are like the P_1 and F_1 plants, but approximately 4 of 1000 are only 12 cm high and about 4 of 1000 are 36 cm high.
(a) What mode of inheritance is occurring here?
(b) How many gene pairs are involved?
(c) How much does each gene contribute to plant height?
(d) Indicate one possible set of genotypes for the original P_1 parents and the F_1 plants that could account for these results.
(e) Indicate three possible genotypes that could account for F_2 plants that are 18 cm high and three that account for F_2 plants that are 33 cm high.

7. Erma and Harvey were a compatible barnyard pair, but a curious sight. Harvey's tail was only 6 cm long, while Erma's was 30 cm. Their F_1 piglet offspring all grew tails that were 18 cm. When inbred, an F_2 generation resulted in many piglets (Erma and Harvey's grandpigs), whose tails ranged in 4-cm intervals from 6 to 30 cm (6, 10, 14, 18, 22, 26, and 30). Most had 18-cm tails, while 1/64 had 6-cm tails and 1/64 had 30-cm tails.
(a) Explain how these tail lengths were inherited by describing the mode of inheritance, indicating how many gene pairs were at work, and designating the genotypes of Harvey, Erma, and their 18-cm-tail offspring.
(b) If one of the 18-cm-tail F_1 pigs is mated with one of the 6-cm-tail F_2 pigs, what phenotypic ratio will be predicted if many offspring resulted? Diagram the cross.

8. In the following table, average differences of height, weight, and fingerprint ridge count between monozygotic twins (reared together and apart), dizygotic twins, and nontwin siblings are compared:

Trait	MZ Reared Together	MZ Reared Apart	DZ Reared Together	Sibs Reared Together
Height (cm)	1.7	1.8	4.4	4.5
Weight (kg)	1.9	4.5	4.5	4.7
Ridge count	0.7	0.6	2.4	2.7

Based on the data in this table, which of these quantitative traits has the highest heritability values?

9. What kind of heritability estimates (broad sense or narrow sense) are obtained from human twin studies?

10. List as many human traits as you can that are likely to be under the control of a polygenic mode of inheritance.

11. Corn plants from a test plot are measured, and the distribution of heights at 10-cm intervals is recorded in the following table:

Height (cm)	Plants (no.)
100	20
110	60
120	90
130	130
140	180
150	120
160	70
170	50
180	40

Calculate (a) the mean height, (b) the variance, (c) the standard deviation, and (d) the standard error of the mean. Plot a rough graph of plant height against frequency. Do the values represent a normal distribution? Based on your calculations, how would you assess the variation within this population?

12. The following variances were calculated for two traits in a herd of hogs.

Trait	V_P	V_G	V_A
Back fat	30.6	12.2	8.44
Body length	52.4	26.4	11.70

(a) Calculate broad-sense (H^2) and narrow-sense (h^2) heritabilities for each trait in this herd.
(b) Which of the two traits will respond best to selection by a breeder? Why?

13. The mean and variance of plant height of two highly inbred strains (P_1 and P_2) and their progeny (F_1 and F_2) are shown here.

Strain	Mean (cm)	Variance
P_1	34.2	4.2
P_2	55.3	3.8
F_1	44.2	5.6
F_2	46.3	10.3

Calculate the broad-sense heritability (H^2) of plant height in this species.

14. A hypothetical study investigated the vitamin A content and the cholesterol content of eggs from a large population of chickens. The following variances (V) were calculated.

	Trait	
Variance	Vitamin A	Cholesterol
V_P	125.5	862.0
V_E	96.2	484.6
V_A	12.0	192.1
V_D	15.3	185.3

(a) Calculate the narrow-sense heritability (h^2) for both traits.
(b) Which trait, if either, is likely to respond to selection?

15. In a herd of dairy cows the narrow-sense heritability for milk protein content is 0.76, and for milk butterfat it is 0.82. The correlation coefficient between milk protein content and butterfat is 0.91. If the farmer selects for cows producing more butterfat in their milk, what will be the most likely effect on milk protein content in the next generation?

16. In an assessment of learning in *Drosophila*, flies were trained to avoid certain olfactory cues. In one population, a mean of 8.5 trials was required. A subgroup of this parental population that was trained most quickly (mean = 6.0) was interbred, and their progeny were examined. These flies demonstrated a mean training value of 7.5. Calculate realized heritability for olfactory learning in *Drosophila*.

17. Suppose you want to develop a population of *Drosophila* that would rapidly learn to avoid certain substances the flies could detect by smell. Based on the heritability estimate you obtained in Problem 16, do you think it would be worth doing this by artificial selection? Why or why not?

18. In a population of tomato plants, mean fruit weight is 60 g and h^2 is 0.3. Predict the mean weight of the progeny if tomato plants whose fruit averaged 80 g were selected from the original population and interbred.

19. In a population of 100 inbred, genotypically identical rice plants, variance for grain yield is 4.67. What is the heritability for yield? Would you advise a rice breeder to improve yield in this strain of rice plants by selection?

20. Many traits of economic or medical significance are determined by quantitative trait loci (QTLs) in which many genes, usually scattered throughout the genome, contribute to expression.
(a) What general procedures are used to identify such loci?
(b) What is meant by the term *cosegregate* in the context of QTL mapping? Why are markers such as RFLPs, SNPs, and microsatellites often used in QTL mapping?

Extra-Spicy Problems

Mastering Genetics Visit for instructor-assigned tutorials and problems.

21. A 3-inch plant was crossed with a 15-inch plant, and all F_1 plants were 9 inches. The F_2 plants exhibited a "normal distribution," with heights of 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15 inches.
(a) What ratio will constitute the "normal distribution" in the F_2?
(b) What will be the outcome if the F_1 plants are testcrossed with plants that are homozygous for all nonadditive alleles?

22. In a cross between a strain of large guinea pigs and a strain of small guinea pigs, the F_1 are phenotypically uniform, with an average size about intermediate between that of the two parental strains. Among 1014 F_2 individuals, 3 are about the same size as the small parental strain and 5 are about the same size as the large parental strain. How many gene pairs are involved in the inheritance of size in these strains of guinea pigs?

23. Type A1B brachydactyly (short middle phalanges) is a genetically determined trait that maps to the short arm of chromosome 5 in humans. If you classify individuals as either having or not having brachydactyly, the trait appears to follow a single-locus, incompletely dominant pattern of inheritance. However, if one examines the fingers and toes of affected individuals, one sees a range of expression from extremely short to only slightly short. What might cause such variation in the expression of brachydactyly?

24. In a series of crosses between two true-breeding strains of peaches, the F_1 generation was uniform, producing 30-g peaches. The F_2 fruit mass ranges from 38 to 22 g at intervals of 2 g.
(a) Using these data, determine the number of polygenic loci involved in the inheritance of peach mass.
(b) Using gene symbols of your choice, give the genotypes of the parents and the F_1.

25. Students in a genetics laboratory began an experiment in an attempt to increase heat tolerance in two strains of *Drosophila melanogaster*. One strain was trapped from the wild six weeks before the experiment was to begin; the other was obtained from a *Drosophila* repository at a university laboratory. In which strain would you expect to see the most rapid and extensive response to heat-tolerance selection, and why?

26. Consider a true-breeding plant, *AABBCC,* crossed with another true-breeding plant, *aabbcc,* whose resulting offspring are *AaBbCc.* If you cross the F_1 generation, and independent assortment is operational, the expected fraction of offspring in each phenotypic class is given by the expression $N!/M!(N - M)!$ where N is the total number of alleles (six in this example) and M is the number of uppercase alleles. In a cross of $AaBbCc \times AaBbCc$, what proportion of the offspring would be expected to contain two uppercase alleles?

27. Canine hip dysplasia is a quantitative trait that continues to affect most large breeds of dogs in spite of approximately 40 years of effort to reduce the impact of this condition. Breeders and veterinarians rely on radiographic and universal registries to facilitate the development of breeding schemes for reducing its incidence. Data [Wood and Lakhani (2003). *Vet. Rec.* 152:69–72] indicate that there is a "month-of-birth" effect on hip dysplasia in Labrador retrievers and Gordon setters, whereby the frequency and extent of expression of this disorder vary depending on the time of year dogs are born. Speculate on how breeders attempt to "select" out this disorder and what the month-of-birth phenomenon indicates about the expression of polygenic traits.

28. Floral traits in plants often play key roles in diversification, in that slight modifications of those traits, if genetically determined, may quickly lead to reproductive restrictions and evolution. Insight into genetic involvement in flower formation is often acquired through selection experiments that expose realized heritability. Lendvai and Levin (2003) conducted a series of artificial selection experiments on flower size (diameter) in *Phlox drummondii*. Data from their selection experiments are presented in the following table in modified form and content.

Year	Treatment	Mean (mm)
1997	Control	30.04
	Selected parents	34.13
	Offspring	32.21
1998	Control	28.11
	Selected parents	31.98
	Offspring	31.90
1999	Control	29.68
	Selected parents	31.81
	Offspring	33.74

(a) Considering that differences in control values represent year-to-year differences in greenhouse conditions, calculate (in mm) the average response to selection over the three-year period.

(b) Calculate the realized heritability for each year and the overall realized heritability.

(c) Assuming that the realized heritability in phlox is relatively high, what factors might account for such a high response?

(d) In terms of evolutionary potential, is a population with high heritability likely to be favored compared to one with a low realized heritability?

29. In 1988, Horst Wilkens investigated blind cavefish, comparing them with members of a sibling species with normal vision that are found in a lake [Wilkens, H. (1988). *Evol. Biol.* 25:271–367]. We will call them cavefish and lakefish. Wilkens found that cavefish eyes are about seven times smaller than lakefish eyes. F_1 hybrids have eyes of intermediate size. These data, as well as the $F_1 \times F_1$ cross and those from backcrosses ($F_1 \times$ cavefish and $F_1 \times$ lakefish), are depicted below. Examine Wilkens's results and respond to the following questions:

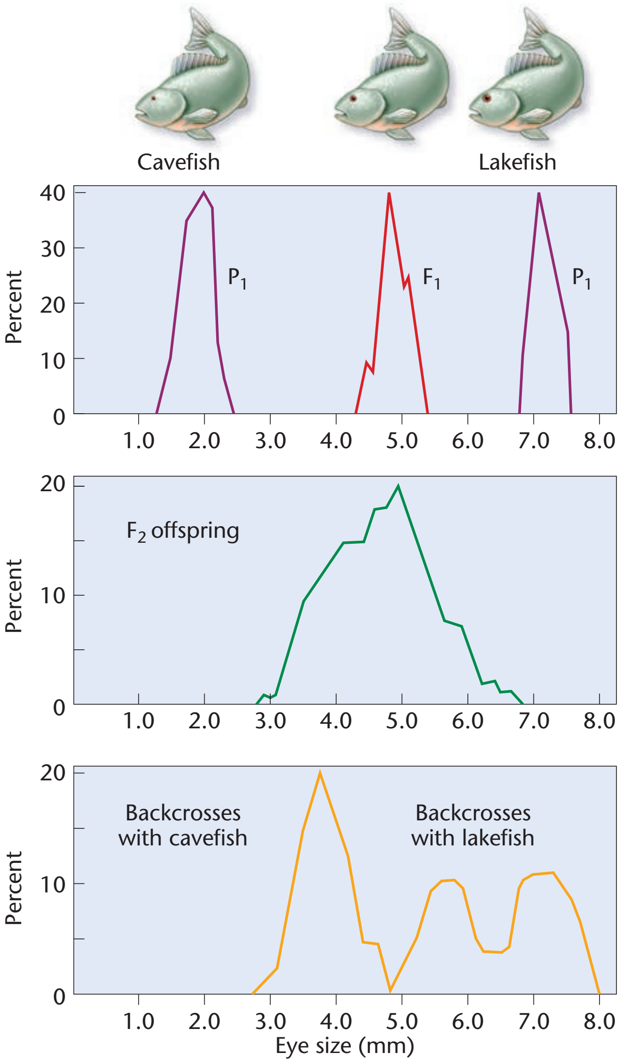
(a) Based strictly on the F_1 and F_2 results of Wilkens's initial crosses, what possible explanation concerning the inheritance of eye size seems most feasible?

(b) Based on the results of the F_1 backcross with cavefish, is your explanation supported? Explain.

(c) Based on the results of the F_1 backcross with lakefish, is your explanation supported? Explain.

(d) Wilkens examined about 1000 F_2 progeny and estimated that 6–7 genes are involved in determining eye size. Is the sample size adequate to justify this conclusion? Propose an experimental protocol to test the hypothesis.

(e) A comparison of the embryonic eye in cavefish and lakefish revealed that both reach approximately 4 mm in diameter. However, lakefish eyes continue to grow, while cavefish eye size is greatly reduced. Speculate on the role of the genes involved in this problem.



26

Population and Evolutionary Genetics

These ladybird beetles, from the Chiricahua Mountains in Arizona, show considerable phenotypic variation.

CHAPTER CONCEPTS

- Most populations and species harbor considerable genetic variation.
- This variation is reflected in the alleles distributed among populations of a species.
- The relationship between allele frequencies and genotype frequencies in an ideal population is described by the Hardy–Weinberg law.
- Selection, migration, and genetic drift can cause changes in allele frequency.
- Mutation creates new alleles in a population gene pool.
- Nonrandom mating changes population genotype frequency but not allele frequency.
- A reduction in gene flow between populations, accompanied by selection or genetic drift, can lead to reproductive isolation and speciation.
- Genetic differences between populations or species are used to reconstruct evolutionary history.

In the mid-nineteenth century, Alfred Russel Wallace and Charles Darwin identified natural selection as the mechanism of evolution. In his book *On the Origin of Species*, published in 1859, Darwin provided evidence that populations and species are not fixed, but change, or evolve, over time due to natural selection. However, Wallace and Darwin could not explain either the origin of the variations that provide the raw material for evolution or the mechanisms by which such variations are passed from parents to offspring. Gregor Mendel published his work on the inheritance of traits in 1866, but it received little notice at the time. After the rediscovery of Mendel's work in 1900, twentieth-century biologists began a 30-year effort to reconcile Mendel's concept of genes and alleles with the theory of evolution by natural selection. As the biologists applied the principles of Mendelian genetics to populations, both the source of variation (mutation and recombination) and the mechanism of inheritance (segregation of alleles) were explained. We now view evolution as a consequence of changes in genetic material through mutation and changes in allele frequencies in populations over time. This union of population genetics with the theory of natural selection generated a new view of the evolutionary process, called *neo-Darwinism*.

In addition to natural selection, other forces including mutation, migration, and drift, individually and collectively, alter allele frequencies and bring about evolutionary divergence that eventually may result in **speciation,** the formation of new species. Speciation is facilitated by environmental diversity. If a population is spread over a geographic range encompassing a number of ecologically distinct subenvironments with different selection pressures, the populations occupying these areas may gradually adapt and become genetically distinct from one another. Genetically differentiated populations may remain in existence, become extinct, reunite with each other, or continue to diverge until they become reproductively isolated. Populations that are reproductively isolated are regarded as separate species. Genetic changes within populations can modify a species over time, transform it into another species, or cause it to split into two or more species.

Population geneticists investigate patterns of genetic variation within and among groups of interbreeding individuals. Changes in genetic structure form the basis for the evolution of a population. Thus, population genetics has become an important subdiscipline of evolutionary biology. In this chapter, we examine the population genetics processes of **microevolution,** which is defined as evolutionary change within populations of a species, and then consider how molecular aspects of these processes can be extended to **macroevolution,** which is defined as evolutionary events leading to the emergence of new species and other taxonomic groups.

26.1 Genetic Variation Is Present in Most Populations and Species

A **population** is a group of individuals belonging to the same species that live in a defined geographic area and actually or potentially interbreed. In thinking about the human population, we can define it as everyone who lives in the United States, or in Sri Lanka, or we can specify a population as all the residents of a particular small town or village.

The genetic information carried by members of a population constitutes that population's **gene pool.** At first glance, it might seem that a population that is well adapted to its environment must have a gene pool that is highly homozygous because it would seem likely that the most favorable allele at each locus is present at a high frequency. In addition, a look at most populations of plants and animals reveals many phenotypic similarities among individuals. However, a large body of evidence indicates that, in reality, most populations contain a high degree of heterozygosity. This built-in genetic variation is not necessarily apparent in the phenotype; hence, detecting it is not a simple task. Nevertheless, the amount of variation within a population can be revealed by several methods.

Detecting Genetic Variation

The detection and use of genetic variation in individuals and populations began long before genetics emerged as a science. Millennia ago, plant and animal breeders began using artificial selection to domesticate plants and animals. However, as genetic technology developed in the last century, the ability to detect and quantify genetic variation in genes, in individual genomes, and in the genomes of populations has grown exponentially.

One of the more spectacular examples of how much variation exists in the gene pool of a species was the use of selective breeding to create hundreds of dog breeds in nineteenth-century England over a period of less than 75 years. Many people, seeing a Chihuahua (about 10 inches high) and a Great Dane (about 42 inches high) for the first time, might find it difficult to believe they are both members of the same species (**Figure 26.1**).

Recombinant DNA Technology and Genetic Variation

After the discovery that DNA carries genetic information and the development of recombinant DNA technology, efforts centered on detecting genetic variation in the sequence of individual genes carried by individuals in a population.

In one such study, Martin Kreitman isolated, cloned, and sequenced copies of the alcohol dehydrogenase (*Adh*) gene from individuals representing five different populations of *Drosophila melanogaster*. The 11 cloned genes from these five populations contained a total of 43 nucleotide differences in the *Adh* sequence of 2721 base pairs (**Figure 26.2**). These variations are distributed throughout the gene: 14 in exon coding regions, 18 within introns, and 11 in untranslated flanking regions. Of the 14 variations in exons, only one leads to an amino acid replacement—the one in codon 192, resulting in the two known alleles of this gene. The other 13 nucleotide changes do not lead to amino acid replacements and are silent variations of this gene.

Genetic Variation in Genomes

The development of **next-generation sequencing technology** has extended the detection of genomic variation

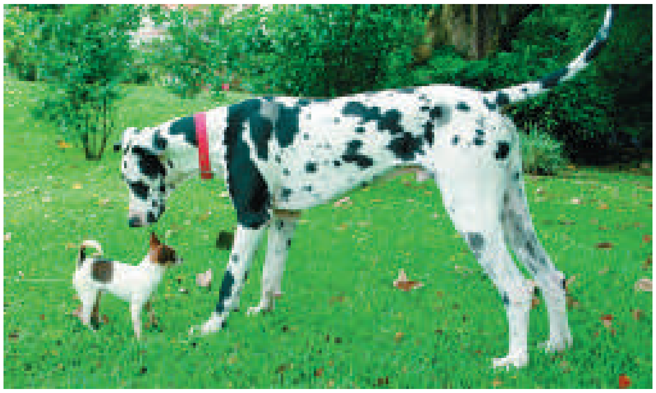


FIGURE 26.1 The size difference between a Chihuahua and a Great Dane illustrates the high degree of genetic variation present in the dog genome.

	Exon 3	Intron 3	Exon 4
Consensus *Adh* sequence:	C C C C	G G A A T	C T C C A*C T A G
Strain			
Wa-S	T T • A	C A • T A	A C • • • • • • •
Fl1-S	T T • A	C A • T A	A C • • • • • • •
Ja-S	• • • •	• • • • •	• • • T • T • C A
Fl-F	• • • •	• • • • •	• • G T C T C C •
Ja-F	• • A •	• • G • •	• • G T C T C C •

FIGURE 26.2 DNA sequence variation in parts of the *Drosophila Adh* gene in a sample of the 11 laboratory strains derived from the five natural populations. The dots represent nucleotides that are the same as the consensus sequence; letters represent nucleotide polymorphisms. An A/C polymorphism (A*) in codon 192 creates the two *Adh* alleles (F and S). All other polymorphisms are silent or noncoding.

from individuals to populations. The 1000 Genomes Project, which ran from 2008 through 2015, was a global effort to identify and catalog at least 95 percent of the common genetic variations carried by the 7 billion people now inhabiting the planet. The Project eventually sequenced the genomes of 2504 individuals from 26 populations using a combination of whole-genome sequencing at low coverage levels, exome sequencing, and microarray genotyping.

Over 88 million genetic variants were identified in the human genome, including 84.7 million SNPs, 3.6 million indels (short insertion and deletions), and 60,000 structural variants [copy number variations (CNVs), Alu and LINE-1 insertions, etc.].

The Project's overall goal is to explore and understand the relationship between genotype and phenotype. In humans, this translates into using association studies to identify variants associated with disease. For example, in studies to date, no single variant has been associated with diabetes; this implies that a combination of heritable multiple rare variants is related to this common disorder. Eventually, researchers hope to associate specific genetic variants with cellular pathways and networks associated with complex disorders such as hypertension, cardiovascular disease, and neurological disorders associated with protein accumulation such as Alzheimer disease and Huntington disease.

Explaining the High Level of Genetic Variation in Populations

The finding that populations harbor considerable genetic variation at the amino acid and nucleotide levels came as a surprise to many evolutionary biologists. The early consensus had been that selection would favor a single optimal (wild-type) allele at each locus and that, as a result, populations would have high levels of homozygosity. This expectation was shown conclusively to be wrong, and considerable research and argument have ensued concerning the forces that maintain such high levels of genetic variation.

The **neutral theory** of molecular evolution, proposed by Motoo Kimura in 1968, proposes that mutations leading to amino acid substitutions are usually detrimental, with only a very small fraction being favorable. Some mutations are neutral; that is, they are functionally equivalent to the allele they replace. Mutations that are favorable or detrimental are preserved or removed from the population, respectively, by natural selection. However, the frequency of the neutral alleles in a population will be determined by mutation rates and random genetic drift, and not by selection. Some neutral mutations will drift to fixation in the population; other neutral mutations will be lost. At any given time, a population may contain several neutral alleles at any particular locus. The diversity of alleles at most loci does not, therefore, reflect the action of natural selection, but instead is a function of population size (larger populations have more variation) and the fraction of mutations that are neutral.

The alternative explanation for the surprisingly high genetic variation in populations is natural selection. There are several extensively documented examples in which enzyme or protein variations are maintained by adaptation to certain environmental conditions. The well-known advantage of sickle-cell anemia heterozygotes when infected by malarial parasites is such an example.

Fitness differences of a fraction of a percent would be sufficient to maintain such a variation, but at that level their presence would be difficult to measure. Current data are therefore insufficient to determine what fraction of molecular genetic variation is neutral and what fraction is subject to selection. The neutral theory nonetheless serves a crucial function: It points out that some genetic variation is expected simply as a result of mutation and drift. In addition, the neutral theory provides a working hypothesis for studies of molecular evolution. In other words, biologists must find positive evidence that selection is acting on allele frequencies at a particular locus before they can reject the simpler assumption that only mutation and drift are at work.

26.2 The Hardy–Weinberg Law Describes Allele Frequencies and Genotype Frequencies in Population Gene Pools

Often when we examine a single gene in a population, we find that different allele combinations of this gene result in individuals with different genotypes. For example, two alleles, A and a, of the A gene can be combined to produce three genotypes: *AA*, *Aa*, and *aa*. Key elements of population genetics depend on the calculation of allele frequencies and genotype frequencies in a gene pool, and the determination of how these frequencies change from one generation to the next. Population geneticists use these calculations to answer questions such as: How much genetic variation is present in a population? Are genotypes randomly distributed in time and space, or do discernible patterns exist? What processes affect the composition of a population's gene pool? Do these processes produce genetic divergence among populations that may lead to the formation of new species? Changes in allele frequencies in a population that do not directly result in species formation are examples of microevolution. In the following sections, we will discuss microevolutionary changes in population gene pools and then will consider macroevolution and the process of speciation.

The relationship between the relative proportions of alleles in the gene pool and the frequencies of different genotypes in the population was elegantly described in a mathematical model developed independently by the British mathematician Godfrey H. Hardy and the German physician Wilhelm Weinberg. This model, called the **Hardy–Weinberg law,** describes what happens to allele and genotype frequencies in an "ideal" population that is infinitely large and randomly mating and that is not subject to any evolutionary forces such as mutation, migration, or selection.

Calculating Genotype Frequencies

The Hardy–Weinberg model uses the principle of Mendelian segregation and simple probability to explain the relationship between allele and genotype frequencies in a population. We can demonstrate how this works by considering a single autosomal gene with two alleles, *A* and *a,* in a population where the frequency of *A* is 0.7 and the frequency of *a* is 0.3. Note that $0.7 + 0.3 = 1$, indicating that all the alleles of gene *A* present in the population are accounted for.

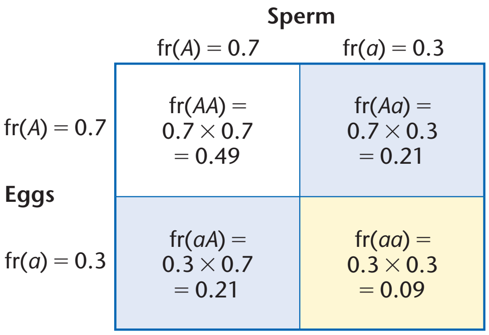
These allele frequencies mean that the probability that any female gamete will contain *A* is 0.7, and the probability that a male gamete will contain *A* is also 0.7. The probability that *both* gametes will contain *A* is $0.7 \times 0.7 = 0.49$. Thus we predict that in the offspring, the genotype *AA* will occur 49 percent of the time. The probability that a zygote will be formed from a female gamete carrying *A* and a male gamete carrying *a* is $0.7 \times 0.3 = 0.21$, and the probability of a female gamete carrying *a* being fertilized by a male gamete carrying *A* is $0.3 \times 0.7 = 0.21$, so the frequency of genotype *Aa* in the offspring is $0.21 + 0.21 = 0.42 = 42$ percent. Finally, the probability that a zygote will be formed from two gametes carrying *a* is $0.3 \times 0.3 = 0.09$, so the frequency of genotype *aa* is 9 percent. As a check on our calculations, note that $0.49 + 0.42 + 0.09 = 1.0$, confirming that we have accounted for all possible genotypic combinations in the zygotes. These calculations are summarized in Figure 26.3.

Calculating Allele Frequencies

Now that we know the frequencies of genotypes in the next generation, what will be the allele frequencies in this new generation? Under the Hardy–Weinberg law, we assume that all genotypes have equal rates of survival and reproduction. This means that in the next generation, all genotypes contribute equally to the new gene pool. The *AA* individuals constitute 49 percent of the population, and we can predict that the gametes they produce will constitute 49 percent of the gene pool. These gametes all carry allele *A*. Similarly, *Aa* individuals constitute 42 percent of the population, so we predict that their gametes will constitute 42 percent of the new gene pool. Half (0.5) of these gametes will carry allele *A*. Thus, the frequency of allele *A* in the gene pool is $0.49 + (0.5)\,0.42 = 0.7$. The other half of the gametes produced by *Aa* individuals will carry allele *a*. The *aa* individuals constitute 9 percent of the population, so their gametes will constitute 9 percent of the new gene pool. All these gametes carry allele *a*. Thus, we can predict that the allele *a* in the new gene pool is $(0.5)\,0.42 + 0.09 = 0.3$. As a check on our calculation, note that $0.7 + 0.3 = 1.0$, accounting for all the gametes in the gene pool of the new generation.

The Hardy–Weinberg Law and Its Assumptions

Because the Hardy–Weinberg law is a mathematical model, we use variables instead of numerical values for the allele frequencies in the general case. Imagine a gene pool in which the frequency of allele *A* is represented by *p*



		Sperm	
		fr(*A*) = 0.7	fr(*a*) = 0.3
Eggs	fr(*A*) = 0.7	fr(*AA*) = 0.7×0.7 = 0.49	fr(*Aa*) = 0.7×0.3 = 0.21
	fr(*a*) = 0.3	fr(*aA*) = 0.3×0.7 = 0.21	fr(*aa*) = 0.3×0.3 = 0.09

FIGURE 26.3 Calculating genotype frequencies from allele frequencies. Gametes represent samples drawn from the gene pool to form the genotypes of the next generation. In this population, the frequency of the *A* allele is 0.7, and the frequency of the *a* allele is 0.3. The frequencies of the genotypes in the next generation are calculated as 0.49 for *AA,* 0.42 for *Aa,* and 0.09 for *aa.* Under the Hardy–Weinberg law, the frequencies of *A* and *a* remain constant from generation to generation.

and the frequency of allele a is represented by q, such that $p + q = 1$. If we randomly draw male and female gametes from the gene pool and pair them to make a zygote, the probability that both will carry allele A is $p \times p$. Thus, the frequency of genotype AA among the zygotes is p^2. The probability that the female gamete carries A and the male gamete carries a is $p \times q$, and the probability that the female gamete carries a and the male gamete carries A is $q \times p$. Thus, the frequency of genotype Aa among the zygotes is $2pq$. Finally, the probability that both gametes carry a is $q \times q$, making the frequency of genotype aa among the zygotes q^2. Therefore, the distribution of genotypes among the zygotes is

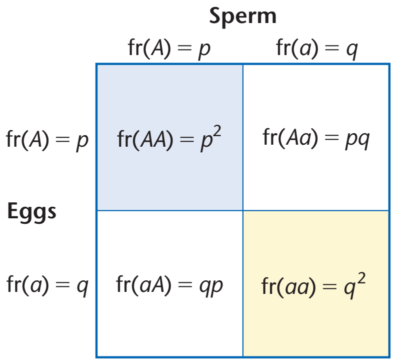
$$p^2 + 2pq + q^2 = 1$$

These calculations are summarized in **Figure 26.4**. They demonstrate the two main predictions of the Hardy–Weinberg model:

1. Allele frequencies in our population do not change from one generation to the next.
2. After one generation of random mating, genotype frequencies can be predicted from the allele frequencies.

In other words, there is no change in allele frequency, and for this locus, the population does not undergo any microevolutionary change. The theoretical population described by the Hardy–Weinberg model is based on the following assumptions:

1. Individuals of all genotypes have equal rates of survival and equal reproductive success—that is, there is no selection.
2. No new alleles are created or converted from one allele into another by mutation.
3. Individuals do not migrate into or out of the population.
4. The population is infinitely large, which in practical terms means that the population is large enough that sampling errors and other random effects are negligible.
5. Individuals in the population mate randomly.



		Sperm	
		fr(A) = p	fr(a) = q
Eggs	fr(A) = p	fr(AA) = p^2	fr(Aa) = pq
	fr(a) = q	fr(aA) = qp	fr(aa) = q^2

FIGURE 26.4 The general description of allele and genotype frequencies under Hardy–Weinberg assumptions. The frequency of allele A is p, and the frequency of allele a is q. After mating, the three genotypes AA, Aa, and aa have the frequencies p^2, $2pq$, and q^2, respectively.

These assumptions are what make the Hardy–Weinberg model so useful in population genetics research. By specifying the conditions under which the population does not evolve, the Hardy–Weinberg model can be used to identify the real-world forces that cause allele frequencies to change. Application of this model can also reveal "neutral genes" in a population gene pool—those not being operated on by the forces of evolution.

The Hardy–Weinberg model has three additional important consequences:

1. Dominant traits do not necessarily increase from one generation to the next.
2. **Genetic variability** can be maintained in a population, since, once established in an ideal population, allele frequencies remain unchanged.
3. Under Hardy–Weinberg assumptions, knowing the frequency of just one genotype enables us to calculate the frequencies of all other genotypes at that locus.

This is particularly useful in human genetics because we can calculate the frequency of heterozygous carriers for recessive genetic disorders even when all we know is the frequency of affected individuals.

Now Solve This

26.1 The ability to taste the compound phenylthiocarbamide (PTC) is controlled by a dominant allele *T*. Individuals homozygous for the recessive allele *t* are unable to taste PTC. In a genetics class of 125 students, 88 can taste PTC and 37 cannot. Calculate the frequency of the *T* and *t* alleles in this population and the frequency of the genotypes.

- **HINT:** *This problem involves an understanding of how to use the Hardy–Weinberg law. The key to its solution is to determine which allele frequency (p or q) you must estimate first when homozygous dominant and heterozygous genotypes have the same phenotype.*

26.3 The Hardy–Weinberg Law Can Be Applied to Human Populations

To show how allele frequencies are measured in a real population, let's consider a gene that influences an individual's susceptibility to infection by HIV-1, the virus responsible for acquired immunodeficiency syndrome (AIDS). A small number of individuals who make high-risk choices (such as having unprotected sex with HIV-positive partners) never become infected. Some of these individuals are homozygous for a mutant allele of a gene called *CCR5*.

The *CCR5* gene (**Figure 26.5**) encodes a protein called the C-C chemokine receptor-5 (CCR5). Chemokines are cell surface signaling molecules associated with the immune system. The CCR5 protein is also used by strains of HIV-1 to gain entry into cells. The mutant allele of the *CCR5* gene contains a 32-bp deletion, making the encoded protein shorter and nonfunctional and blocking the entry of HIV-1 into cells. The normal allele is called *CCR51* (also called *1*), and the mutant allele is called *CCR5-Δ32* (also called Δ*32*).

Individuals homozygous for the mutant allele (*Δ32/Δ32)* are resistant to HIV-1 infection. Heterozygous (*1/Δ32)* individuals are susceptible to HIV-1 infection but progress more slowly to AIDS. **Table 26.1** summarizes the genotypes possible at the *CCR5* locus and the phenotypes associated with each.

The discovery of the *CCR5-Δ32* allele generates two important questions: Which human populations carry this allele, and how common is it? To address these questions, teams of researchers surveyed members of several populations. Genotypes were determined by direct analysis of DNA (**Figure 26.6**). In one population, 79 individuals had genotype *1/1*, 20 had genotype *1/Δ32*, and 1 individual had genotype *Δ32/Δ32*. We can see that this population had 158 *1* alleles carried by the *1/1* individuals plus 20 *1* alleles carried by the *1/Δ32* individuals, for a total of 178. The frequency of the *CCR51* allele in the sample population is thus 178/200 = 0.89 = 89 percent. Twenty *1/Δ32* individuals and two *Δ32/Δ32* individuals each carried a copy of the *CCR5-Δ32* allele, for a total of 22. The frequency of the *CCR5-Δ32* allele is thus 22/200 = 0.11 = 11 percent. Notice that $p + q = 1$, confirming that we have accounted for all the alleles of the *CCR51* gene in the population. **Table 26.2** shows two methods for computing the frequencies of the alleles in the population surveyed.

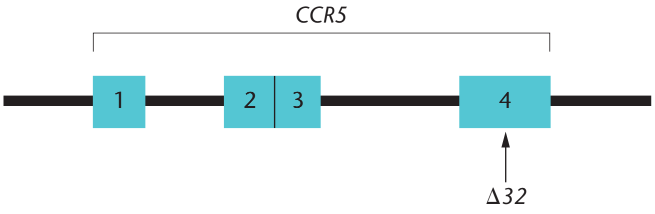


FIGURE 26.5 The organization of the *CCR5* gene in region 3p21.3 of human chromosome 3. The gene contains 4 exons and 2 introns (there is no intron between exons 2 and 3). The arrow shows the location of the 32-bp deletion in exon 4 that confers resistance to HIV-1 infection.

TABLE 26.1 ***CCR5*** **Genotypes and Phenotypes**

Genotype	Phenotype
1/1	Susceptible to sexually transmitted strains of HIV-1
1/Δ32	Susceptible but may progress to AIDS slowly
Δ32/Δ32	Resistant to most sexually transmitted strains of HIV-1

Can we expect the *CCR5-Δ32* allele to increase in human populations because it offers resistance to infection by HIV? This specific question is difficult to answer directly, but as we will see later in this chapter, when factors such as natural selection, mutation, migration, or genetic drift are present, the allele frequencies in a population may change from one generation to the next.

By determining allele frequencies over more than one generation, it is possible to determine whether the frequencies remain in equilibrium because the Hardy–Weinberg assumptions are operating. Populations that meet the Hardy–Weinberg assumptions are not evolving because allele frequencies (for the generations tested) are not changing. However, a population may be in Hardy–Weinberg equilibrium for the alleles being tested, but other genes may not be in equilibrium.

Testing for Hardy–Weinberg Equilibrium in a Population

One way to see if any of the Hardy–Weinberg assumptions do not hold in a given population is to determine whether the population's genotypes are in equilibrium.

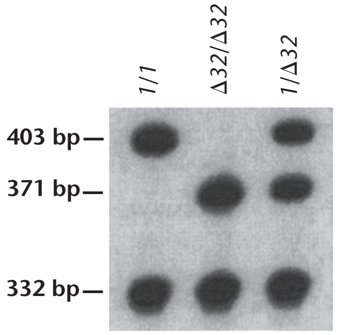


FIGURE 26.6 Allelic variation in the *CCR5* gene. Michel Samson and colleagues used polymerase chain reaction (PCR) to amplify a part of the *CCR5* gene containing the site of the 32-bp deletion, cut the resulting DNA fragments with a restriction enzyme, and ran the fragments on an electrophoresis gel. Each lane reveals the genotype of a single individual. The *1* allele produces a 332-bp fragment and a 403-bp fragment; the *Δ32* allele produces a 332-bp fragment and a 371-bp fragment. Heterozygotes produce three bands.

TABLE 26.2 Methods of Determining Allele Frequencies from Data on Genotypes

	Genotype			
(a) Counting Alleles	***1/1***	***1/Δ32***	***Δ32/Δ32***	**Total**
Number of individuals	79	20	1	100
Number of *1* alleles	158	20	0	178
Number of *Δ32* alleles	0	20	2	22
Total number of alleles	158	40	2	200
Frequency of *CCR5-1* in sample: 178/200 = 0.89 = 89%				
Frequency of *CCR5*-Δ32 in sample: 22/200 = 0.11 = 11%				

	Genotype			
(b) From Genotype Frequencies	***1/1***	***1/Δ32***	***Δ32/Δ32***	**Total**
Number of individuals	79	20	1	100
Genotype frequency	79/100 = 0.79	20/100 = 0.20	1/100 = 0.01	1.00
Frequency of CCR5-1 in sample: 0.79 + (0.5)0.20 = 0.89 = 89%				
Frequency of *CCR5-Δ32* in sample: (0.5)0.20 + 0.01 = 0.11 = 11%				

To do this, we first determine the genotype frequencies. This can be done directly from the phenotypes (if heterozygotes are recognizable), by analyzing proteins or DNA sequences, or indirectly, using the frequency of the HIV-1 resistant phenotype in the population to calculate genotype frequencies using the Hardy—Weinberg law. We can then calculate the allele frequencies from the genotype frequencies. Finally, the allele frequencies in the parental generation are used to predict the genotype frequencies in the next generation. According to the Hardy—Weinberg law, genotype frequencies are predicted to fit the $p^2 + 2pq + q^2 = 1$ relationship. If they do not, then one or more of the assumptions are invalid for the population in question.

To demonstrate, let's examine *CCR5* genotypes in a hypothetical population. Our population is composed of 283 individuals; of these, 223 have genotype *1/1*; 57 have genotype *1/Δ32;* and 3 have genotype *Δ32/Δ32.* These numbers represent the following genotype frequencies: *1/1* = 223/283 = 0.788, 1/*Δ32* = 57/283 = 0.201, and *Δ32/Δ32* = 3/283 = 0.011, respectively. From the genotype frequencies, we can compute the *CCR5-1* allele frequency as 0.89 and the *CCR5-Δ32* allele frequency as 0.11. Once we know the allele frequencies, we can use the Hardy—Weinberg law to determine whether this population is in equilibrium. The allele frequencies predict the genotype frequencies in the next generation as follows:

- Expected frequency of genotype *1/1*:

$$p^2 = (0.89)^2 = 0.792$$

- Expected frequency of genotype *1/Δ32:*

$$2pq = 2(0.89)(0.11) = 0.196$$

- Expected frequency of genotype *Δ32/Δ32:*

$$q^2 = (0.11)^2 = 0.012$$

These expected frequencies are nearly identical to the frequencies observed in the parental generation. Our test of this population has failed to provide evidence that Hardy—Weinberg assumptions are being violated. The conclusion can be confirmed by using the whole numbers utilized in calculating the genotype frequencies to perform a χ^2 analysis (see Chapter 3). In this case, neither the genotype frequencies nor the allele frequencies are changing in this population, meaning that the population is in equilibrium. As we will see in later sections of this chapter, forces such as natural selection, mutation, migration, and chance operate to bring about changes in allele frequency. These forces drive both microevolution and the formation of new species (macroevolution).

Now Solve This

26.2 Determine whether the following two sets of data represent populations that are in Hardy–Weinberg equilibrium.

(a) *CCR5* genotypes: *1/1*, 60 percent; *1/Δ32*, 35.1 percent; *Δ32/Δ32,* 4.9 percent

(b) Sickle-cell hemoglobin: *SS,* 75.6 percent; *Ss,* 24.2 percent; *ss,* 0.2 percent (*S* = normal hemoglobin allele; *s* = mutant hemoglobin allele)

HINT: *This problem involves an understanding of how to use the Hardy—Weinberg law to determine whether populations are in genetic equilibrium. The key to its solution is to first determine the allele frequencies based on the genotype frequencies provided.*

Calculating Frequencies for Multiple Alleles in Populations

Although we have used one-gene, two-allele systems as examples, many genes have several alleles, all of which can be found in a single population. The ABO blood group in humans (discussed in Chapter 4) is such an example. The locus *I* (isoagglutinin) has three alleles, I^A, I^B, and *i*, yielding six possible genotypic combinations (I^AI^A, I^BI^B, *ii*, I^AI^B, I^Ai, I^Bi). Remember that in this case I^A and I^B are codominant alleles and that both of these are dominant to *i*. The result is that homozygous I^AI^A and heterozygous I^Ai individuals are phenotypically identical, as are I^BI^B and I^Bi individuals, so we can distinguish only four phenotypic blood-type combinations: Type A, Type B, Type AB, and Type O.

By adding another variable to the Hardy–Weinberg equation, we can calculate both the genotype and allele frequencies for the situation involving three alleles. Let *p, q,* and *r* represent the frequencies of alleles I^A, I^B, and *i*, respectively. Note that because there are three alleles

$$p + q + r = 1$$

Under Hardy–Weinberg assumptions, the frequencies of the genotypes are given by

$$(p + q + r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$$

If we know the frequencies of blood types for a population, we can then estimate the frequencies for the three alleles of the ABO system. For example, in one population sampled, the following blood-type frequencies are observed: A = 0.53, B = 0.133, O = 0.26. Because the *i* allele is recessive, the population's frequency of Type O blood equals the proportion of the recessive genotype r^2. Thus,

$$\begin{aligned} r^2 &= 0.26 \\ r &= \sqrt{0.26} \\ &= 0.51 \end{aligned}$$

Using *r,* we can calculate the allele frequencies for the I^A and I^B alleles. The I^A allele is present in two genotypes, I^AI^A and I^Ai. The frequency of the I^AI^A genotype is represented by p^2 and the I^Ai genotype by $2pr$. Therefore, the combined frequency of Type A blood and Type O blood is given by

$$p^2 + 2pr + r^2 = 0.53 + 0.26$$

If we factor the left side of the equation and take the sum of the terms on the right,

$$\begin{aligned} (p + r)^2 &= 0.79 \\ p + r &= \sqrt{0.79} \\ p &= 0.89 - r \\ &= 0.89 - 0.51 = 0.38 \end{aligned}$$

Having calculated *p* and *r,* the frequencies of allele I^A and allele *i*, we can now calculate the frequency for the I^B allele:

$$\begin{aligned} p + q + r &= 1 \\ q &= 1 - p - r \\ &= 1 - 0.38 - 0.51 \\ &= 0.11 \end{aligned}$$

The phenotypic and genotypic frequencies for this population are summarized in **Table 26.3**.

Calculating Allele Frequencies for X-linked Traits

The Hardy–Weinberg law can be used to calculate allele and genotype frequencies for X-linked traits, as long as we remember that in an XY sex-determination system, the homogametic (XX) sex has two copies of an X-linked allele, whereas the heterogametic sex (XY) only has one copy. Thus, for mammals (including humans) where the female is XX and the male is XY, the frequency of the X-linked allele in the gene pool and the frequency of males expressing the X-linked trait will be the same. This is because each male only has one X chromosome, and the probability that any individual male receives an X chromosome with the allele in question must be equal to the frequency of the allele. The probability of any individual female having the allele in question on both X chromosomes will be q^2, where *q* is the frequency of the allele.

TABLE 26.3 **Calculating Genotype Frequencies for Multiple Alleles in a Hardy–Weinberg Population Where the Frequency of Allele $I^A = 0.38$, Allele $I^B = 0.11$, and Allele $i = 0.51$**

Genotype	Genotype Frequency	Phenotype	Phenotype Frequency
I^AI^A	$p^2 = (0.38)^2 = 0.14$	A	0.53
I^Ai	$2pr = 2(0.38)(0.51) = 0.39$		
I^BI^B	$q^2 = (0.11)^2 = 0.01$	B	0.12
I^Bi	$2qr = 2(0.11)(0.51) = 0.11$		
I^AI^B	$2pr = 2(0.38)(0.11) = 0.084$	AB	0.08
ii	$r^2 = (0.51)^2 = 0.26$	O	0.26

To illustrate this for a recessive X-linked trait, let's consider the example of red-green color blindness, which affects 8 percent of human males. The frequency of the color blindness allele is therefore 0.08; in other words, 8 percent of X chromosomes carry it.

The other 92 percent of X chromosomes carry the dominant allele for normal red-green color vision. If we define p as the frequency of the normal allele and q as the frequency of the color blindness allele, then $p = 0.92$ and $q = 0.08$. The frequency of color-blind females (with two affected X chromosomes) is $q^2 = (0.08)^2 = 0.0064$, and the frequency of carrier females (having one normal and one affected X chromosome) is $2pq = 2(0.08)(0.92) = 0.147$. In other words, 14.7 percent of females carry the allele for red-green color blindness and can pass it to their children, although they themselves have normal color vision.

An important consequence of the difference in allele frequency for X-linked genes between male and female gametes is that for a rare recessive allele, the trait will be expressed at a much higher frequency among XY individuals than among those who are XX. So, for example, diseases such as hemophilia and Duchenne muscular dystrophy (DMD) in humans, both of which are caused by recessive mutations on the X chromosome, are much more common in boys, who need only to inherit a single copy of the mutated allele to suffer from the disease. Girls who inherit two affected X chromosomes will also have the disease; but with a rare allele, the probability of this occurrence is very small.

Calculating Heterozygote Frequency

A useful application of the Hardy–Weinberg law, especially in human genetics, allows us to estimate the frequency of heterozygotes in a population. To do this, we must first calculate the frequency of each allele in the population. Although homozygous unaffected and heterozygous individuals have the same phenotype, we can usually determine the frequency of a recessive trait by first identifying and counting individuals with the homozygous recessive phenotype in a sample of the population. Using this information and the Hardy–Weinberg law, we can then calculate both the allele and genotype frequencies for each genotype present in a population.

Cystic fibrosis, an autosomal recessive trait, has an incidence of about 1/2500 (0.0004) in people of northern European ancestry. Individuals with cystic fibrosis are easily distinguished from the population at large by such symptoms as extra-salty sweat, excess amounts of thick mucus in the lungs, and susceptibility to bacterial infections. Because this is a recessive trait, individuals with cystic fibrosis must be homozygous. Their frequency in a population is represented by q^2 (provided that mating has been random in the previous generation). The frequency of the recessive allele is therefore

$$q = \sqrt{q^2} = \sqrt{0.0004} = 0.02$$

Knowing that the frequency of the recessive allele is about 2 percent, we can calculate the frequency of the normal (dominant) allele because $p + q = 1$. Using this equation, the frequency of p is

$$p = 1 - q = 1 - 0.02 = 0.98$$

Now that the allele frequencies are known, we can calculate the frequency of the heterozygous genotype. In the Hardy–Weinberg equation, the frequency of heterozygotes is $2pq$. Thus,

$$\begin{aligned} 2pq &= 2(0.98)(0.02) \\ &= 0.04 = 4\% = 1/25 \end{aligned}$$

The results show that heterozygotes for cystic fibrosis are rather common (about 1/25 individuals, or 4 percent of the population), even though the frequency of homozygous recessives is only 1/2500, or 0.04 percent. However, keep in mind that this calculation of heterozygote frequency is an estimate because the population in question may not meet all Hardy–Weinberg assumptions.

In general, for a single locus with a dominant and recessive allele, the frequencies of all three genotypes (homozygous dominant, heterozygous, homozygous recessive) can be estimated once the frequency of either allele is known and Hardy–Weinberg assumptions are invoked. The relationship between genotype and allele frequency is shown in **Figure 26.7**. It is important to note that heterozygotes increase rapidly in a population as the values of p and q move from 0 or 1.0 towards 0.5. This observation confirms our conclusion that when a recessive trait such as cystic fibrosis is rare, the majority of those carrying the allele are

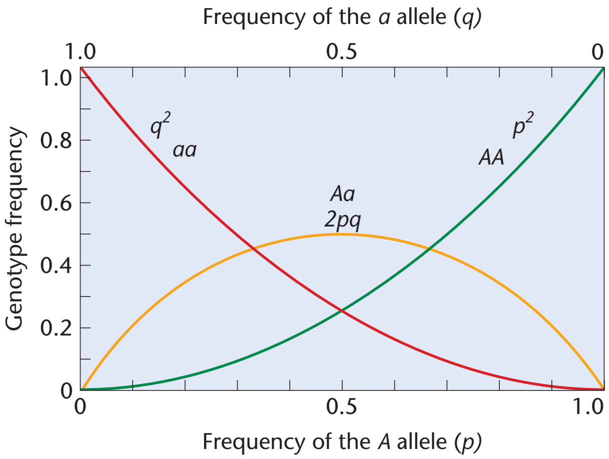


FIGURE 26.7 The relationship between genotype and allele frequencies derived from the Hardy–Weinberg equation.

heterozygotes. In populations in which the frequencies of p and q are between 0.33 and 0.67, heterozygotes occur at a higher frequency than either homozygote.

Now Solve This

26.3 If the albino phenotype occurs in 1/10,000 individuals in a population at equilibrium and albinism is caused by an autosomal recessive allele *a,* calculate the frequency of:

(a) The recessive mutant allele
(b) The normal dominant allele
(c) Heterozygotes in the population
(d) Mating between heterozygotes

- **HINT:** *This problem involves an understanding of the method of calculating allele and genotype frequencies. The key to its solution is to first determine the frequency of the albinism allele in this population.*

26.4 Natural Selection Is a Major Force Driving Allele Frequency Change

To understand evolution, we must understand the forces that transform the gene pools of populations and can lead to the formation of new species. Chief among the mechanisms transforming populations is **natural selection,** discovered independently by Alfred Russel Wallace and Charles Darwin. The Wallace–Darwin concept of natural selection can be summarized as follows:

1. Individuals of a species exhibit variations in phenotype, for example, differences in size, agility, coloration, defenses against enemies, ability to obtain food, courtship behaviors, and flowering times.
2. Many of these variations, even small and seemingly insignificant ones, are heritable and are passed on to offspring.
3. Organisms tend to reproduce in an exponential fashion. More offspring are produced than can survive. This causes members of a species to engage in a struggle for survival, competing with other members of the community for scarce resources. Offspring also must avoid predators, and in sexually reproducing species, adults must compete for mates.
4. In the struggle for survival, individuals with particular phenotypes will be more successful than others, allowing the former to survive and reproduce at higher rates.

As a consequence of natural selection, populations and species change. Traits that promote differential survival and reproduction will become more common, and traits that confer a lowered ability for survival and reproduction will become less common. This means that over many generations, traits that confer a reproductive advantage will increase in frequency, which in turn will cause the population to become better adapted to its current environment. Over time, if selection continues, it may result in the appearance of new species.

Detecting Natural Selection in Populations

Recall that measuring allele frequencies and genotype frequencies using the Hardy–Weinberg law is based on several assumptions about an ideal population: large population size, lack of migration, presence of random mating, absence of selection and mutation, and equal survival rates of offspring.

However, if all genotypes do not have equal rates of survival or do not leave equal numbers of offspring, then allele frequencies may change from one generation to the next. To see why, let's imagine a population of 100 individuals in which the frequency of allele A is 0.5 and that of allele a is 0.5. Assuming the previous generation mated randomly, we find that the genotype frequencies in the present generation are $(0.5)^2 = 0.25$ for *AA,* $2(0.5)(0.5) = 0.5$ for *Aa,* and $(0.5)^2 = 0.25$ for *aa.* Because our population contains 100 individuals, we have 25 *AA* individuals, 50 *Aa* individuals, and 25 *aa* individuals.

Now let's suppose that individuals with different genotypes have different rates of survival: All 25 *AA* individuals survive to reproduce, 90 percent or 45/50 of the *Aa* individuals survive to reproduce, and 80 percent or 20/25 of the *aa* individuals survive to reproduce. When the survivors reproduce, each contributes two gametes to the new gene pool, giving us $2(25) + 2(45) + 2(20) = 180$ gametes. What are the frequencies of the two alleles in the surviving population? We have 50 A gametes from *AA* individuals, plus 45 A gametes from *Aa* individuals, so the frequency of allele A is $(50 + 45)/180 = 0.53$. We have 45 a gametes from *Aa* individuals, plus 40 a gametes from *aa* individuals, so the frequency of allele a is $(45 + 40)/180 = 0.47$.

These differ from the frequencies we started with. The frequency of allele A has increased, whereas the frequency of allele a has declined. A difference in survival or reproduction rate (or both) among individuals is an example of natural selection, which is the principal force that shifts allele frequencies within large populations. Natural selection is one of the most important factors in evolutionary change.

Fitness and Selection

Selection occurs whenever individuals with a particular genotype enjoy an advantage in survival or reproduction over other genotypes. However, selection may vary over a wide range, from much less than 1 percent to 100 percent.

In the previous hypothetical example, selection was strong. Weak selection might involve just a fraction of a percent difference in the survival rates of different genotypes. Advantages in survival and reproduction ultimately translate into increased genetic contribution to future generations. An individual organism's genetic contribution to future generations is called its **fitness.** Genotypes associated with high rates of reproductive success are said to have high fitness, whereas genotypes associated with low rates of reproductive success are said to have low fitness.

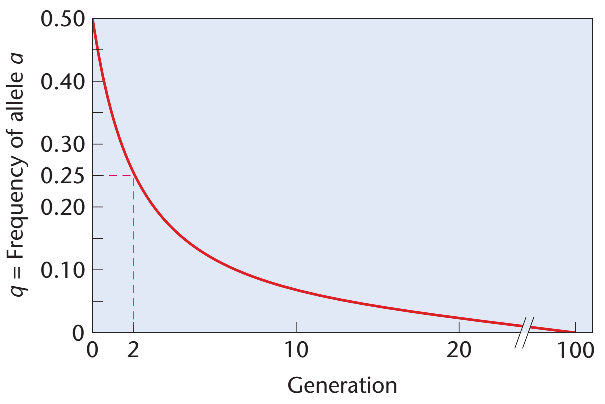
Hardy–Weinberg analysis also allows us to examine fitness as a measure of the degree of natural selection. By convention, population geneticists use the letter w to represent fitness. Thus, w_{AA} represents the relative fitness of genotype *AA*, w_{Aa} the relative fitness of genotype *Aa,* and w_{aa} the relative fitness of genotype *aa.* Assigning the values $w_{AA} = 1, w_{Aa} = 0.9$, and $w_{aa} = 0.8$ would mean, for example, that all *AA* individuals survive, 90 percent of *Aa* individuals survive, and 80 percent of *aa* individuals survive, as in the previous hypothetical case.

Let's consider selection against deleterious alleles. Fitness values $w_{AA} = 1, w_{Aa} = 1$, and $w_{aa} = 0$ describe a situation in which *a* is a homozygous lethal allele. As homozygous recessive individuals die without leaving offspring, the frequency of allele *a* will decline. The decline in the frequency of allele *a* is described by the equation

$$q_g = \frac{q_0}{1 + gq_0}$$

where q_g is the frequency of allele *a* in generation g, q_0 is the starting frequency of *a* (i.e., the frequency of *a* in generation zero), and *g* is the number of generations that have passed.

Figure 26.8 shows what happens to a lethal recessive allele with an initial frequency of 0.5. At first, because of the high percentage of *aa* genotypes, the frequency of allele *a* declines rapidly. The frequency of *a* is halved in only two generations. By the sixth generation, the frequency is halved again. By now, however, the majority of *a* alleles are carried by heterozygotes. Because *a* is recessive, these heterozygotes are not selected against. Consequently, as more time passes, the frequency of allele *a* declines ever more slowly. As long as heterozygotes continue to mate, it is difficult for selection to completely eliminate a recessive allele from a population.



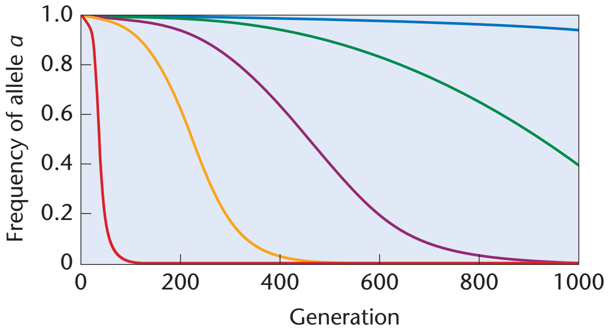
Generation	p	q	p^2	$2pq$	q^2
0	0.50	0.50	0.25	0.50	0.25
1	0.67	0.33	0.44	0.44	0.12
2	0.75	0.25	0.56	0.38	0.06
3	0.80	0.20	0.64	0.32	0.04
4	0.83	0.17	0.69	0.28	0.03
5	0.86	0.14	0.73	0.25	0.02
6	0.88	0.12	0.77	0.21	0.01
10	0.91	0.09	0.84	0.15	0.01
20	0.95	0.05	0.91	0.09	< 0.01
40	0.98	0.02	0.95	0.05	< 0.01
70	0.99	0.01	0.98	0.02	< 0.01
100	0.99	0.01	0.98	0.02	< 0.01

FIGURE 26.8 The change in the frequency of a lethal recessive allele, *a*. The frequency of *a* is halved in two generations and halved again by the sixth generation. Subsequent reductions occur slowly because the majority of *a* alleles are carried by heterozygotes.

Figure 26.9 shows the outcome of different degrees of selection against a nonlethal recessive allele, *a*. In this case, the intensity of selection varies from strong (red curve) to weak (blue curve), as well as intermediate values (yellow, purple, and green curves). In each example, the frequency of the deleterious allele, *a,* starts at 0.99 and declines over time. However, the rate of decline depends heavily on the strength of selection. When selection is strong and only 90 percent of the heterozygotes and 80 percent of the *aa* homozygotes survive (red curve), the frequency of allele *a* drops from 0.99 to less than 0.01 in about 85 generations. However, when selection is weak, and 99.8 percent of the heterozygotes and 99.6 percent of the *aa* homozygotes survive (blue curve), it takes 1000 generations for the frequency of allele *a* to drop from 0.99 to 0.93. Two important conclusions can be drawn from this example. First, over thousands of generations, even weak selection can cause substantial changes in allele frequencies; because evolution generally occurs over a large number of generations, selection is a powerful force in evolutionary change. Second, for selection to produce rapid changes in allele frequencies, the differences in fitness among genotypes must be large.

There Are Several Types of Selection

The phenotype is the result of the combined influence of the individual's genotype at many different loci and the effects of the environment. Selection can be classified as (1) directional, (2) stabilizing, or (3) disruptive.



Selection Against Allele *a*					
	Strong ←				→ Weak
	🟥	🟧	🟪	🟦	🟦
w_{AA}	1.0	1.0	1.0	1.0	1.0
w_{Aa}	0.90	0.98	0.99	0.995	0.998
w_{aa}	0.80	0.96	0.98	0.99	0.996

FIGURE 26.9 The effect of selection on allele frequency. The rate at which a deleterious allele is removed from a population depends heavily on the strength of selection.

In **directional selection** traits at one end of a spectrum of phenotypes present in the population become selected for or against, usually as a result of changes in the environment. A carefully documented example comes from research by Peter and Rosemary Grant and their colleagues, who study the medium ground finches (*Geospiza fortis*) of Daphne Major Island in the Galápagos Islands. These researchers discovered that the beak size of these birds varies over time in response to fluctuations in the environment (**Figure 26.10**). In 1977, a severe drought killed some 80 percent of the finches on the island. Big-beaked birds survived at higher rates than small-beaked birds because when food became scarce, the big-beaked birds were able to eat a greater variety of seeds, especially larger ones with hard shells. After the drought ended, more plants were available, selection was relaxed, and beak size declined. Droughts in 1980 and 1982 again saw differential survival and reproduction, again shifting the average beak size toward one phenotypic extreme representing larger beak size.

Stabilizing selection, in contrast, selects for intermediate phenotypes, with those at both extremes being selected against. Over time, this will reduce the phenotypic variance in the population but without a significant shift in the mean. One of the clearest demonstrations of stabilizing selection is from a study of human birth weight and survival for 13,730 children born over an 11-year period. **Figure 26.11** shows the distribution of birth weight, the percentage of mortality at five weeks, and the percentage of births in the population (at right). Infant mortality increases on either side of the optimal birth weight of 7.5 pounds. Stabilizing selection acts to keep a population well adapted to its current environment.

Disruptive selection is selection against intermediate phenotypes and selection for phenotypes at both extremes. It can be viewed as the opposite of stabilizing selection because the intermediate types are selected against. This will result in a population with an increasingly bimodal distribution for a trait, as we can see in **Figure 26.12**. In experiments using *Drosophila*, after several generations of disruptive artificial selection for bristle number, in which only flies with high- or low-bristle numbers were allowed to breed, most flies could be easily placed in a low- or high-bristle category. In natural populations, such a situation might exist for a population in a heterogeneous environment.

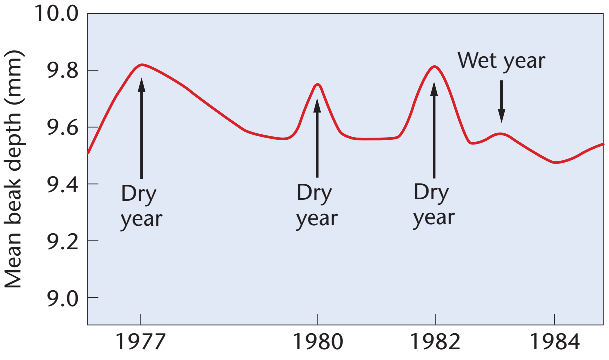


FIGURE 26.10 Beak size in finches during dry years increases because of strong selection. Between droughts, selection for large beak size is not as strong, and birds with smaller beak sizes survive and reproduce, increasing the number of birds with smaller beaks.

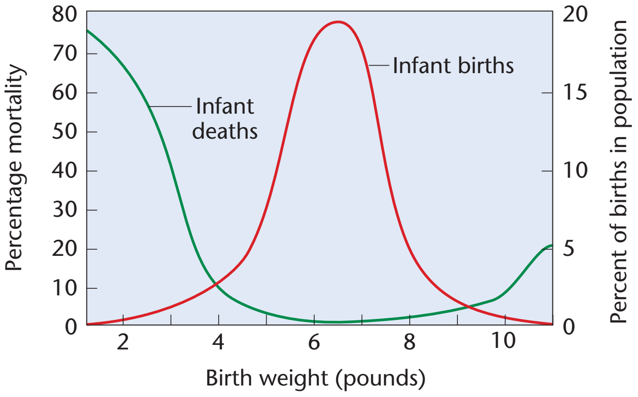


FIGURE 26.11 The relationship between birth weight and mortality in humans.

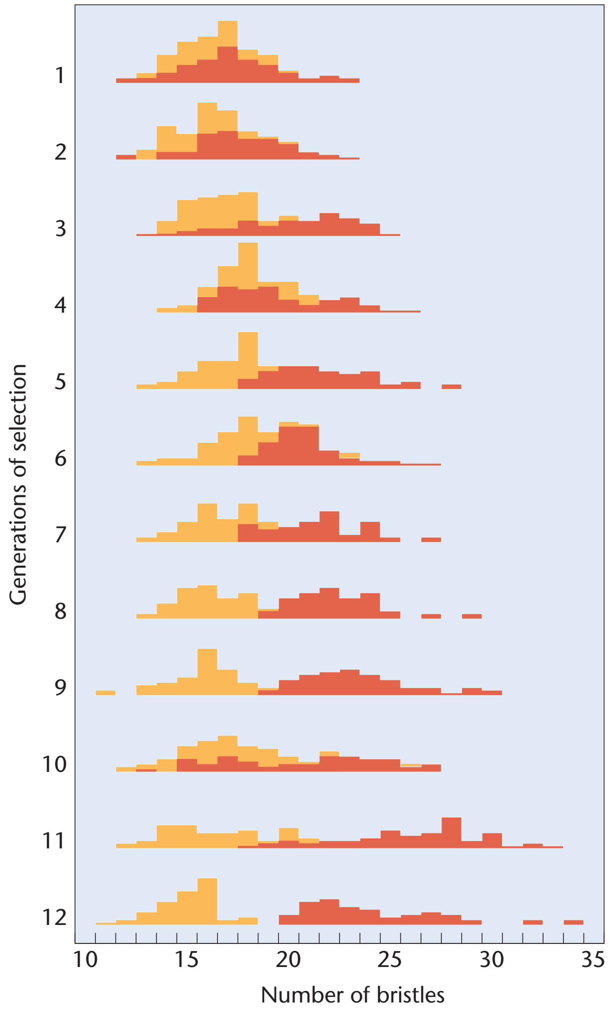


FIGURE 26.12 The effect of disruptive selection on bristle number in *Drosophila*. When individuals with the highest and lowest bristle numbers were selected, the population showed a nonoverlapping divergence in only 12 generations.

The effects of these three forms of selection can be compared by considering their effects on the phenotypic mean and the amount of phenotypic variation present in a population (**Figure 26.13**). In directional selection, one phenotypic extreme is selected for. This causes an increase in the frequency of the favored allele as a result of differences in fitness (survival and reproduction) among the different phenotypes. This shift occurs independently of whether the allele in question is dominant or recessive. As a result, over time, the population mean shifts in the direction of one extreme phenotype. Directional selection allows for rapid changes in allele frequency and is an important factor in speciation.

In stabilizing selection, rather than selecting for one or the other extreme phenotype, both phenotypic extremes are selected against. This negative selection results in increased fitness of the intermediate phenotype and a reduced level of variability, with little or no effect on the mean. Such selection favors maximum adaptation to the existing environment but reduces the phenotypic and genetic diversity of the population.

Disruptive selection is bidirectional and favors both phenotypic extremes while selecting against intermediate phenotypes. This form of selection changes both the mean values and the phenotypic variance in the population. As disruptive selection proceeds, the number of individuals with intermediate phenotypes decreases, the total variance increases, and two distinct subpopulations form, each with its own mean value.

In conclusion, although each form of selection acts differentially on the phenotypic mean and variance of a population, they each play an important role in speciation by altering allele frequency.

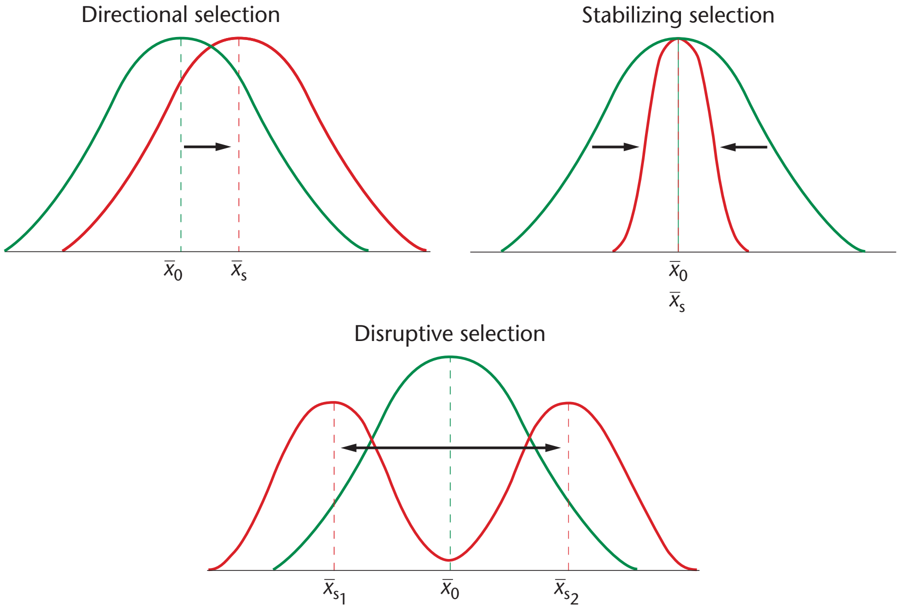


FIGURE 26.13 The impact of directional, stabilizing, and disruptive selection on phenotypic mean and variance. In each case, the mean of the original population $\bar{x}_0$ (green) and the mean of the population following selection $\bar{x}_s$ (red) is shown, along with changes in the amount of phenotypic variance.

26.5 Mutation Creates New Alleles in a Gene Pool

Within a population, the gene pool is reshuffled each generation to produce new offspring. The enormous genetic variation present in the gene pool allows assortment and recombination to produce new combinations of genes already present in the gene pool. But assortment and recombination do not produce new alleles. **Mutation** alone acts to create new alleles. It is important to keep in mind that mutational events occur at random—that is, without regard for any possible benefit or disadvantage to the organism. Mutations not only create new alleles, but in very small populations they can change allele frequencies. Let's consider a population of 20 individuals, and a gene with two alleles, *A* and *a*. If the frequency of *A* is 0.90, the frequency of *a* is 0.10. A mutational event changes one *A* allele into an *a* allele. This event reduces the frequency of the *A* allele from 0.90 to 0.85 and increases the frequency of the *a* allele from 0.10 to 0.15. In this section, we consider whether mutation, by itself, in the larger case, is a significant factor in changing allele frequencies.

To determine whether mutation is a significant force in changing allele frequencies, we measure the rate at which mutations are produced. As in our example, most mutations are recessive, so it is difficult to observe mutation rates directly in diploid organisms. Indirect methods use probability and statistics or large-scale screening programs to estimate mutation rates. For certain dominant mutations, however, a direct method of measurement can be used. To ensure accuracy, several conditions must be met:

1. The allele must produce a distinctive phenotype that can be distinguished from similar phenotypes produced by recessive alleles.
2. The trait must be fully expressed or completely penetrant so that mutant individuals can be identified.
3. An identical phenotype must never be produced by nongenetic agents such as drugs or chemicals.

Suppose for a given gene that undergoes mutation to a dominant allele, 2 out of 100,000 births exhibit a mutant phenotype, but the parents are phenotypically normal. Because the zygotes that produced these births each carry two copies of the gene, we have actually surveyed 200,000 copies of the gene (or 200,000 gametes). If we assume that the affected births are each heterozygous, we have uncovered 2 mutant alleles out of 200,000. Thus, the mutation rate is 2/200,000 or 1/100,000, which in scientific notation is written as 1×10^{-5}. In humans, a dominant form of dwarfism known as **achondroplasia** fulfills the requirements for measuring mutation rates. Individuals with this skeletal disorder have an enlarged skull and short arms and legs. They can be diagnosed by X-ray examination at birth. In a survey of almost 250,000 births, the mutation rate μ for achondroplasia has been calculated as

$$\mu = 1.4 \times 10^{-5} \pm 0.5 \times 10^{-5}$$

Knowing the rate of mutation, we can estimate the extent to which mutation can change allele frequencies from one generation to the next. We represent the normal allele as *d* and the allele for achondroplasia as *D*.

Instead of a population of 20 individuals, imagine a population of 500,000 individuals in which everyone has genotype *dd*. The initial frequency of *d* is 1.0, and the initial frequency of *D* is 0. If each individual contributes two gametes to the gene pool, the gene pool will contain 1,000,000 gametes, all carrying allele *d*. Although the gametes are in the gene pool, 1.4 of every 100,000 *d* alleles mutate into a *D* allele. The frequency of allele *d* is now $(1{,}000{,}000 - 14)/1{,}000{,}000 = 0.999986$, and the frequency of allele *D* is $14/1{,}000{,}000 = 0.000014$. From these numbers, it will clearly be a long time before mutation, by itself, causes any appreciable change in the allele frequencies in this population. In other words, mutation generates new alleles but, unless the population is very small, by itself does not alter allele frequencies at an appreciable rate.

26.6 Migration and Gene Flow Can Alter Allele Frequencies

The Hardy–Weinberg law assumes that migration does not take place. However, occasionally, **migration,** or gene flow, occurs when individuals move between populations. Migration reduces the genetic differences between populations of a species and can increase the level of genetic variation in some populations.

Imagine a species in which a given locus has two alleles, *A* and *a*. There are two populations of this species, one on a mainland and one on an island. The frequency of *A* on the mainland is represented by p_m, and the frequency of *A* on the island is p_i. If there is migration from the mainland to the island, the frequency of *A* in the next generation on the island $p_i{}'$ is given by

$$p_i{}' = (1 - m)p_i + mp_m$$

where m represents migrants from the mainland to the island and that migration is random with respect to genotype.

As an example of how migration might affect the frequency of A in the next generation on the island p_i', assume that $p_i = 0.4$ and $p_m = 0.6$ and that 10 percent of the parents of the next generation are migrants from the mainland ($m = 0.1$). In the next generation, the frequency of allele A on the island will therefore be

$$\begin{aligned} p_i' &= [(1 - 0.1) \times 0.4] + (0.1 \times 0.6) \\ &= 0.36 + 0.06 \\ &= 0.42 \end{aligned}$$

In this case, the flow of genes from the mainland has changed the frequency of A on the island from 0.40 to 0.42 in a single generation.

These calculations reveal that the change in allele frequency attributable to migration is proportional to the differences in allele frequency between the donor and recipient populations *and* to the rate of migration. If either m is large or p_m is very different from p_i, then a rather large change in the frequency of A can occur in a single generation. If migration is the only force acting to change the allele frequency on the island, then equilibrium will be attained when $p_i = p_m$. These guidelines can often be used to estimate migration in cases where it is difficult to quantify. Even in large populations, over time, the effect of migration can substantially alter allele frequencies in populations, as shown for the I^B allele of the ABO blood group in **Figure 26.14**.

26.7 Genetic Drift Causes Random Changes in Allele Frequency in Small Populations

In small populations, significant random fluctuations in allele frequencies are possible by chance alone, a situation known as **genetic drift.** In addition to small population size, drift can arise through the **founder effect,** which occurs when a population originates from a small number of individuals. Although the population may later increase to a large size, the genes carried by all members are derived from those of the founders (assuming no mutation, migration, or selection, and the presence of random mating). Drift can also arise via a **genetic bottleneck.** Bottlenecks develop when a large population undergoes a drastic but temporary reduction in numbers. Even though the population recovers, its genetic diversity has been greatly reduced. In summary, drift is a product of chance and can arise through small population size, founder effects, and bottlenecks. In the following section, we will examine how founder effects can affect allele frequencies.

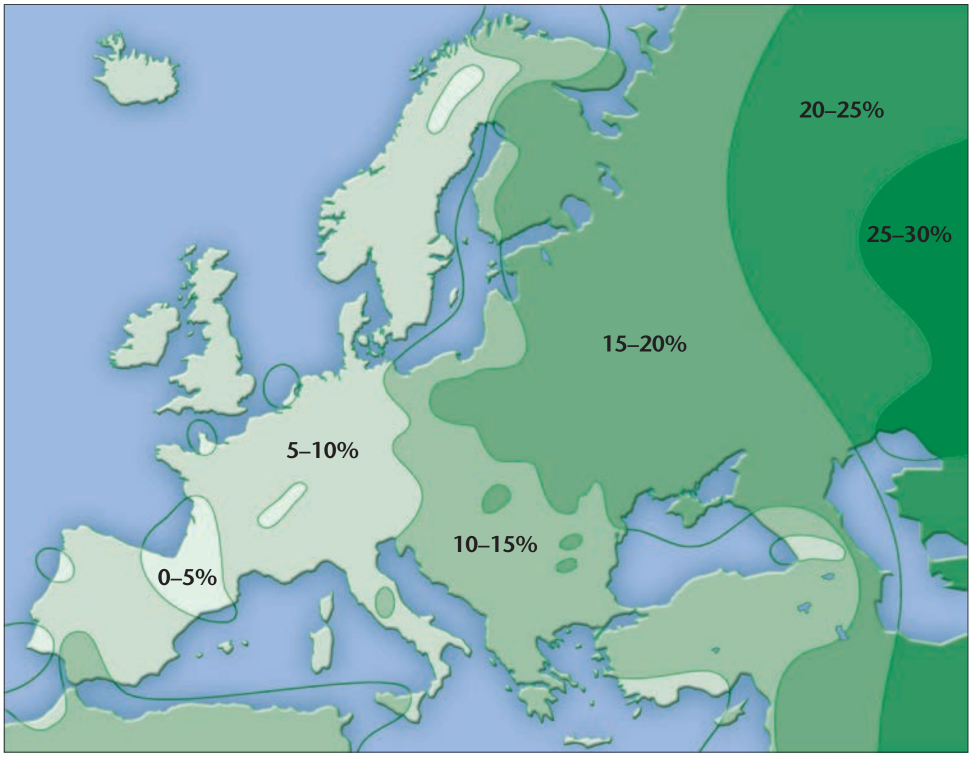


FIGURE 26.14 Migration as a force in evolution. The I^Ballele of the *ABO* locus is present in a gradient from east to west. This allele shows the highest frequency in central Asia and the lowest in northeastern Spain. The gradient parallels the waves of Mongol migration into Europe following the fall of the Roman Empire and is a genetic relic of human history.

Founder Effects in Human Populations

Allele frequencies in certain human populations demonstrate the role of genetic drift in natural populations. Native Americans living in the southwestern United States have a high frequency of oculocutaneous albinism (OCA). In the Navajo, who live primarily in northeastern Arizona, albinism occurs with a frequency of 1 in 1500–2000, compared with whites (1 in 36,000) and African-Americans (1 in 10,000). There are four different forms of OCA (OCA1–4), all with varying degrees of melanin deficiency in the skin, eyes, and hair. OCA2 is caused by mutations in the *P* gene, which encodes a plasma membrane protein. To investigate the genetic basis of albinism in the Navajo, researchers screened for mutations in the *P* gene. In their study, all Navajo with albinism were homozygous for a 122.5-kb deletion in the *P* gene, spanning exons 10–20. This deletion allele was not present in 34 individuals belonging to other Native American populations.

Using a set of PCR primers spanning the deletion, researchers were able to identify homozygous affected individuals, heterozygous carriers, and homozygous normal individuals (**Figure 26.15**). They surveyed 134 normally pigmented Navajo and 42 members of the Apache, a tribe closely related to the Navajo. Based on this sample, the heterozygote frequency in the Navajo is estimated to be 4.5 percent. No carriers were found in the Apache population that was studied.

The 122.5-kb deletion allele causing OCA2 albinism was found only in the Navajo population and not in members of other Native American tribes in the southwestern United States, suggesting that the mutant allele is specific to the Navajo and may have arisen in a single individual who was one of the small number of founders of the Navajo population. Workers originally estimated the age of the mutation to be between 400 and 11,000 years, but tribal history and Navajo oral tradition indicated that the Navajo and Apache became separate populations between 600 and 1000 years ago. Because the deletion is not found in the Apaches, it probably arose in the Navajo population after the tribes split. On this basis, the deletion is estimated to be 400–1000 years old and probably arose as a founder mutation.

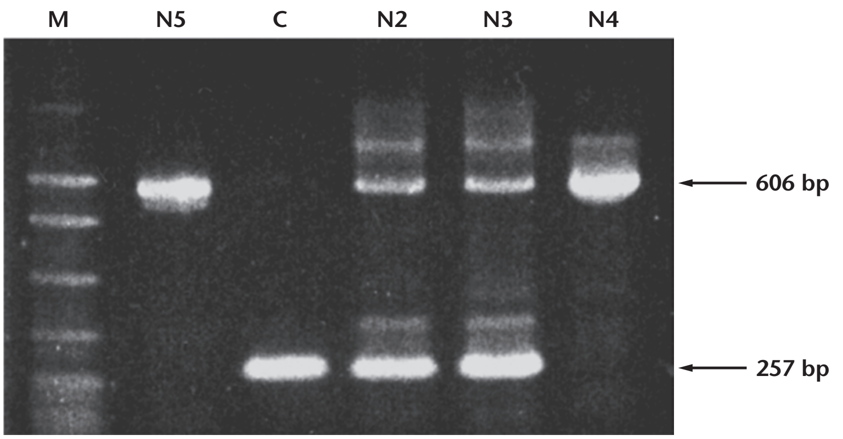


FIGURE 26.15 PCR screens of Navajo individuals affected with albinism (N4 and N5) and the parents of N4 (N2 and N3). Affected individuals (N4 and N5) have a single dense band at 606 bp; heterozygous carriers (N2 and N3) have two bands, one at 606 bp and one at 257 bp.The homozygous normal individual (C) has a single dense band at 257 bp. Each genotype produces a distinctive band pattern, allowing detection of heterozygous carriers in the population. Molecular size markers (M) are in the first lane.

26.8 Nonrandom Mating Changes Genotype Frequency but Not Allele Frequency

We have explored how populations that do not meet the first four assumptions of the Hardy–Weinberg law, in the form of selection, mutation, migration, and genetic drift, can have changes in allele frequencies. The fifth assumption is that members of a population mate at random; in other words, any one genotype has an equal probability of mating with any other genotype in the population. Nonrandom mating can change the frequencies of genotypes in a population. Subsequent selection for or against certain genotypes has the potential to affect the overall frequencies of the alleles they contain, but it is important to note that nonrandom mating *does not itself directly change allele frequencies*.

Nonrandom mating can take one of several forms. In **positive assortative mating,** similar genotypes are more likely to mate than dissimilar ones. This often occurs in humans: A number of studies have indicated that many people are more attracted to individuals who physically resemble them (and are therefore more likely to be genetically similar as well). **Negative assortative mating** occurs when dissimilar genotypes are more likely to mate; some plant species have inbuilt recognition systems that prevent fertilization between individuals with the same alleles at key loci. However, the form of nonrandom mating most commonly found to affect genotype frequencies in population genetics is **inbreeding.**

Inbreeding

Inbreeding occurs when mating individuals are more closely related than any two individuals drawn from the population at random; loosely defined, inbreeding is mating among relatives. For a given allele, inbreeding increases the proportion of homozygotes and decreases the proportion of heterozygotes in the population. A completely inbred population will theoretically consist only of homozygous genotypes. A high level of inbreeding can be harmful because it increases the probability that the number of individuals homozygous for deleterious and/or lethal alleles will increase in the population.

To describe the intensity of inbreeding in a population, Sewall Wright devised the **coefficient of inbreeding (*F*).** This coefficient quantifies the probability that the two alleles of a given gene present in an individual are identical *because they are descended from the same single copy of the allele in an ancestor*. If $F = 1$, all individuals in the population are homozygous, and both alleles in every individual are derived from the same ancestral copy. If $F = 0$, no individual has two alleles derived from a common ancestral copy.

One simple method of estimating *F* for a population is based on the inverse relationship between inbreeding and the frequency of heterozygotes: As the level of inbreeding increases, the frequency of heterozygotes declines. Therefore, *F* can be calculated as

$$F = \frac{H_e - H_o}{H_e}$$

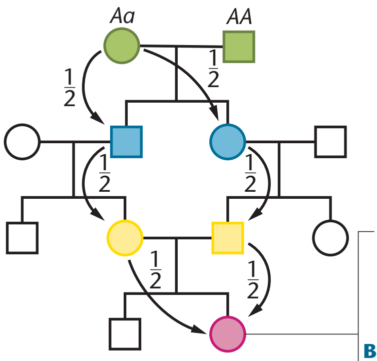
where H_e is the expected heterozygosity based on the Hardy–Weinberg equation and H_o is the observed heterozygosity in the population. Note that if mating in the population is completely at random, the expected and observed levels of heterozygosity will be equal and $F = 0$. A different method that can be used for estimating *F* for an individual is shown in **Figure 26.16**. The fourth-generation female (shaded pink) is the daughter of first cousins (yellow). Suppose her great-grandmother (green) was a carrier of a recessive lethal allele, *a*. What is the probability that the fourth-generation female will inherit two copies of her great-grandmother's lethal allele? For this to happen, (1) the great-grandmother had to pass a copy of the allele to her son, (2) her son had to pass it to his daughter, and (3) his daughter had to pass it to her daughter (the pink female). Also, (4) the great-grandmother had to pass a copy of the allele to her daughter, (5) her daughter had to pass it to her son, and (6) her son had to pass it to his daughter (the pink female). Each of the six necessary events has an individual probability of 1/2, and they *all* have to happen, so the probability that the pink female will inherit two copies of her great-grandmother's lethal allele is $(1/2)^6 = 1/64$. However, to calculate an overall value of the inbreeding coefficient *F* for the pink female as a child of a first-cousin marriage, remember that she could also inherit two copies of any of the other three dominant alleles present in her great-grandparents. Because any of four possibilities would give the pink female two alleles identical by descent from an ancestral copy, although not necessarily two copies of the lethal *a* allele,

$$F = 4 \times (1/64) = 1/16$$

Now Solve This

26.4 A prospective groom, who is unaffected, has a sister with cystic fibrosis (CF), an autosomal recessive disease. Their parents are normal. The brother plans to marry a woman who has no history of CF in her family. What is the probability that they will produce a CF child? They are both Caucasian, and the overall frequency of CF in the Caucasian population is 1/2500—that is, 1 affected child per 2500. (Assume the population meets the Hardy–Weinberg assumptions.)

- **HINT:** *This problem involves an understanding of how recessive alleles are transmitted (see Chapter 3) and the probability of receiving a recessive allele from a heterozygous parent. The key to its solution is to first work out the probability that each parent carries the mutant allele.*



The chance that this female will inherit two copies of her great-grandmother's *a* allele is

$$F = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{64}$$

Because the female's two alleles could be identical by descent from any of four different alleles,

$$F = 4 \times \frac{1}{64} = \frac{1}{16}$$

FIGURE 26.16 Calculating the coefficient of inbreeding *F* for the offspring of a first-cousin marriage.

26.9 Speciation Can Occur through Reproductive Isolation

A **species** can be defined as a group of actually or potentially interbreeding organisms that is reproductively isolated in nature from all other such groups. In sexually reproducing organisms, speciation transforms the parental species into another species or divides a single species into two or more separate species (**Figure 26.17**). Changes in morphology or physiology and adaptations to ecological niches may also occur but are not necessary components of the speciation event.

Populations within a species may carry considerable genetic variation, present as differences in alleles or allele frequencies at a variety of loci. Genetic divergence of these populations that result in different allele frequencies and/or different alleles in their gene pools can reflect the action of forces such as natural selection, mutation, and genetic drift.

When gene flow between populations is reduced or absent, the populations may diverge to the point that members of one population are no longer able to interbreed successfully with members of the other. When populations reach the point where they are reproductively isolated from one another, they have become different species. The genetic changes that result in reproductive isolation between or among populations and lead to the formation of new species or higher taxonomic groups define the process of macroevolution.

The biological barriers that prevent or reduce interbreeding between populations are called **reproductive isolating mechanisms** (**Table 26.4**). These mechanisms may be ecological, behavioral, seasonal, mechanical, or physiological.

Prezygotic isolating mechanisms prevent individuals from mating in the first place. Individuals from different populations may not find each other at the right time, may not recognize each other as suitable mates, or may try to mate but find that they are unable to do so because of differences in mating behavior.

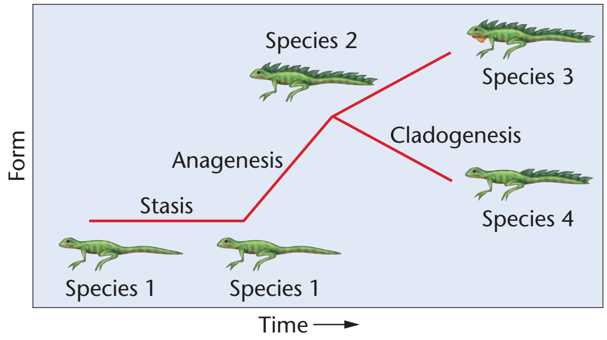


FIGURE 26.17 After a period with no change (stasis), species 1 is transformed into species 2, a process called *anagenesis*. Later, species 2 splits into two new species (species 3 and 4), a process called *cladogenesis*.

Postzygotic isolating mechanisms create reproductive isolation even when the members of two populations are willing and able to mate with each other. For example, mating may take place, and hybrid zygotes may be formed, but all or most of them may be inviable. Alternatively, the hybrids may be viable, but be sterile or suffer from reduced fertility. Yet again, the hybrids themselves may be fertile, but their progeny may have lowered viability or fertility. In all these situations, hybrids are genetic dead-ends. These postzygotic mechanisms act at or beyond the level of the zygote and are generated by genetic divergence.

Postzygotic isolating mechanisms waste gametes and zygotes and lower the reproductive fitness of hybrid survivors. Selection will therefore favor the spread of alleles that lead to the development of prezygotic isolating mechanisms, which in turn prevent interbreeding and the formation of hybrid zygotes and offspring. In animal evolution, one of the most effective prezygotic mechanisms is behavioral isolation, involving courtship behavior.

Changes Leading to Speciation

One form of speciation depends on the formation of geographic barriers between populations, which prevents gene flow between the isolated populations. Isolation allows the gene pools of these populations to diverge.

TABLE 26.4 Reproductive Isolating Mechanisms

Prezygotic Mechanisms: Prevent fertilization and zygote formation.

1. **Geographic or ecological**. The populations live in the same regions but occupy different habitats.
2. **Seasonal or temporal**. The populations live in the same regions but are sexually mature at different times.
3. **Behavioral.** (Only in animals.) The populations are isolated by different and incompatible behavior before mating.
4. **Mechanical**. Cross-fertilization is prevented or restricted by differences in reproductive structures (genitalia in animals, flowers in plants).
5. **Physiological**. Gametes fail to survive in alien reproductive tracts.

Postzygotic Mechanisms: Fertilization takes place and hybrid zygotes are formed, but these are nonviable or give rise to weak or sterile hybrids.

1. **Hybrid nonviability or weakness.**
2. **Developmental hybrid sterility**. Hybrids are sterile because gonads develop abnormally or meiosis breaks down before completion.
3. **Segregational hybrid sterility**. Hybrids are sterile because of abnormal segregation into gametes of whole chromosomes, chromosome segments, or combinations of genes.

If the isolated populations later come into contact, several outcomes are possible. If reproductive isolating mechanisms are not in place, members of these populations will mate and will be regarded as one species. However, if reproductive isolating mechanisms have developed, the two populations will be regarded as separate species.

Formation of the Isthmus of Panama about 3 million years ago created a land bridge connecting North and South America and separated the Caribbean Sea from the Pacific Ocean. After identifying seven Caribbean species of snapping shrimp (**Figure 26.18**) and seven similar Pacific species, researchers matched them in pairs. Analysis of allele frequencies and mitochondrial DNA sequences confirmed that the ancestors of each pair were members of a single species. When the isthmus closed, each of the seven ancestral species was divided into two separate, isolated populations, one in the Caribbean and the other in the Pacific. But after 3 million years of separation, were members of these populations different species?

Males and females were paired together, and successful matings between Caribbean—Pacific couples versus those of Caribbean—Caribbean or Pacific—Pacific pairs were determined. In three of the seven species pairs, transoceanic couples refused to mate altogether. Of the transoceanic pairs that mated, only 1 percent produced viable offspring, while 60 percent of same-ocean pairs produced viable offspring. We can conclude that 3 million years of separation has resulted in complete or nearly complete speciation, involving strong pre- and postzygotic isolating mechanisms for all seven species pairs.

FIGURE 26.18 A snapping shrimp (genus *Alpheus*).

The Rate of Macroevolution and Speciation

How much time is required for speciation? As we saw in the previous example, the time needed for genetic divergence and formation of new species can occur over a span of several million years. In fact, the average time for speciation ranges from 100,000 to 10,000,000 years. However, rapid speciation over much shorter time spans has been reported in a number of cases, including fishes in East African lakes, marine salmon, palm trees on isolated islands, polyploid plants, and brown algae in the Baltic Sea.

In Nicaragua, Lake Apoyo was formed within the last 23,000 years in the crater of a volcano (**Figure 26.19**). This small lake is home to two species of cichlid fish: the Midas cichlid, *Amphilophus citrinellus*, and the Arrow cichlid, *A. zaliosus*. The Midas is the most common cichlid in the region and is found in nearby lakes; the Arrow cichlid is found only in Lake Apoyo.

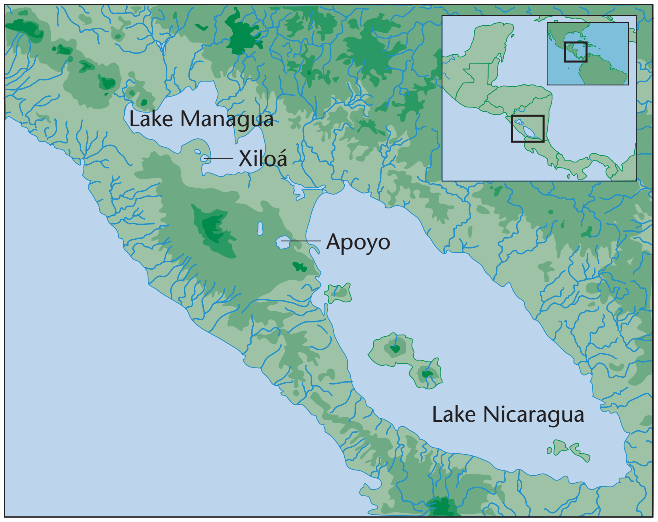


FIGURE 26.19 Lake Apoyo in Nicaragua occupies the crater of an inactive volcano. The lake formed about 23,000 years ago. Two species of cichlid fish in the lake share a close evolutionary relationship.

To establish the evolutionary origin of the Arrow cichlid, researchers used a variety of approaches, including phylogenetic, morphological, and ecological analyses. Sequence analysis of mitochondrial DNA established that the two species form a group with a common ancestor (a monophyletic group). Further genomic analysis of both species using a PCR-based method strengthened the conclusion that these two species are monophyletic and that *A. zaliosus* evolved from *A. citrinellus*. Members of the two species have distinctive morphologies (**Figure 26.20**), including jaw specializations that reflect different food preferences, which were confirmed by analysis of stomach contents. In addition, the two species are reproductively isolated, a conclusion substantiated by laboratory experiments. Using a molecular clock calibrated for cichlid mtDNA, researchers have estimated that *A. zaliosus* evolved from *A. citrinellus* sometime within the last 10,000 years. This estimate, and examples from other species, provides unambiguous evidence that, depending on the strength of selection and that of other parameters of the Hardy—Weinberg law, species formation can occur over a much shorter time scale than the usual range of 100,000—10,000,000 years.

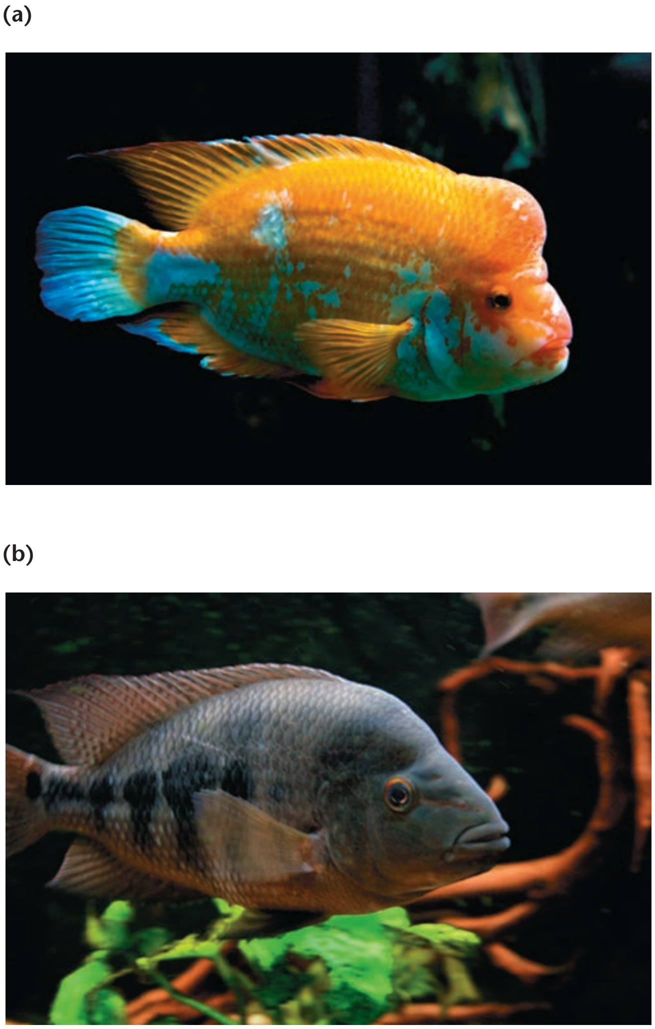


FIGURE 26.20 The two species of cichlids in Lake Apoyo exhibit distinctive morphologies: (a) *Amphilophus citrinellus*. (b) *Amphilophus zaliosus*.

26.10 Phylogeny Can Be Used to Analyze Evolutionary History

Speciation is associated with genetic divergence of populations. Therefore, we should be able to use genetic differences and similarities among present-day species to reconstruct their evolutionary histories. These relationships are most often presented in the form of phylogenetic trees (**Figure 26.21**), which show the ancestral relationships among a group of organisms. These groups can be species, or larger groups such as phyla. In a phylogenetic tree, branches represent the relationships among lineages over time. The length of a branch can be derived from a time scale, showing the length of time between speciation events. Branch points, or nodes, show when a species split into two or more species. Each node represents a common ancestor of the species diverging at that node. The tips of the branches represent species (or a larger group) alive today (or those that ended in extinction). Groups that consist of an ancestral species and all its descendants are called monophyletic groups. The root of a phylogenetic tree represents the oldest common ancestor to all the groups shown in the tree. Trees can be constructed from differences in morphology of living organisms; fossils; and the molecular sequences of proteins, RNA, and DNA.

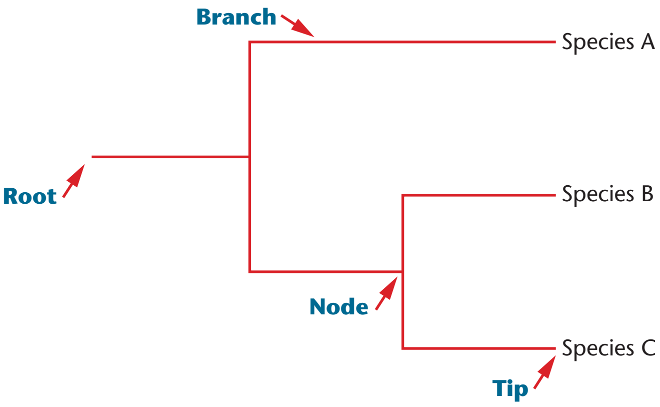


FIGURE 26.21 Elements of a phylogenetic tree showing the relationships among several species. The root represents a common ancestor to all species on the tree. Branches represent lineages through time. The points at which the branches separate are called nodes, and at the tips of the branches are the living or extinct species.

Constructing Phylogenetic Trees from DNA Sequences

Advances in DNA-sequencing technology have made genetic and genomic information from many species available, and today, most phylogenetic trees are constructed using DNA sequences.

Constructing a species-level phylogenetic tree using DNA sequences involves three steps:

1. DNA sequences representing a gene or genome of interest from a number of different species must be acquired. With the proliferation of DNA-sequencing projects, these are usually available from public databases.

2. The sequences must be aligned with each other so that the related parts of each sequence can be compared to see if they are the same or different. The sequences to be compared can be imported into software programs that maximize the number of aligned base pairs by inserting gaps as needed. As discussed earlier, more distantly related species have acquired more DNA differences because of the longer time that has elapsed since they last shared a common ancestor. More closely related species have fewer DNA differences because there has been less time for accumulation of DNA differences since they last shared a common ancestor.

3. These DNA differences are used to construct a phylogenetic tree, often beginning with the most closely related sequences and working backward through sequences that are less closely related.

Reconstructing Vertebrate Evolution by Phylogenetic Analysis

One of the most important steps in the evolutionary history of our species was the ancient transition of vertebrates from the ocean to the land. For more than a century, biologists have debated and argued about which group of lobe-finned fish crawled ashore as the ancestor of all terrestrial vertebrates (amphibians, reptiles, birds, and mammals). In past years, phylogenetic trees constructed from the fossil record, from living species, and from mitochondrial DNA sequences pointed to the lungfish [**Figure 26.22(a)**] as the closest living relative to terrestrial vertebrates, but could not rule out the possibility that vertebrates may have two common ancestors, the lungfish and another fish, the coelacanth [**Figure 26.22(b)**].

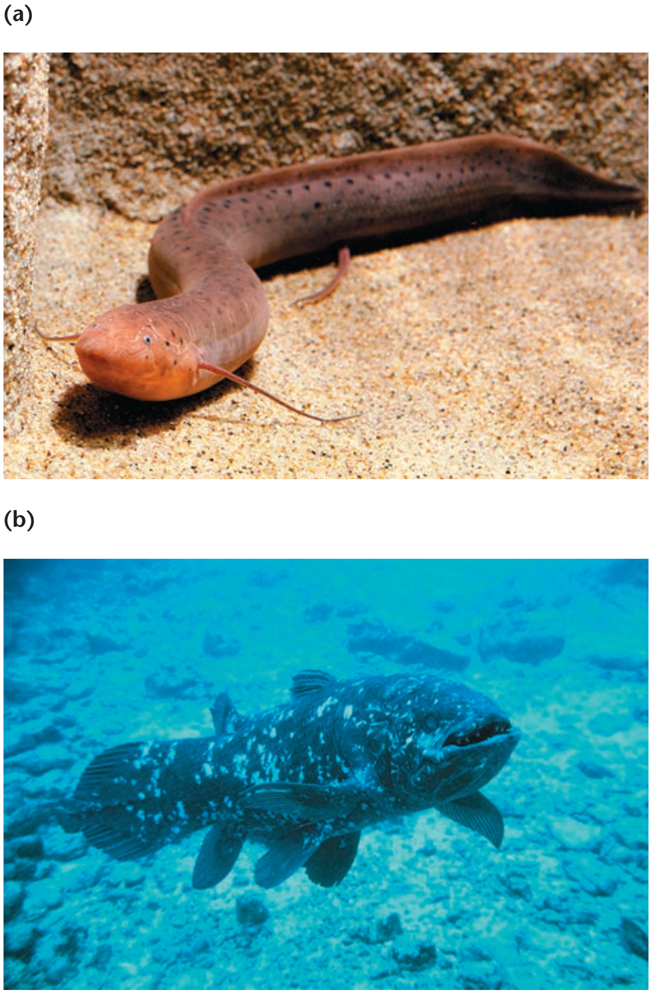


FIGURE 26.22 Phylogenetic evidence indicates that the lungfish (a) and not the coelacanth (b) is a common ancestor of amphibians, reptiles, birds, and mammals.

Recently, the coelacanth genome has been sequenced, and the data from this study have reopened the question of which group shares a common ancestor with our species and all other land vertebrates. Using sequence data from the coelacanth, the lungfish, and selected vertebrate species, researchers aligned and analyzed information from 251 protein-coding genes to construct a phylogenetic tree (**Figure 26.23**). The results strongly support earlier work indicating that terrestrial vertebrates are more closely related to the lungfish than to the coelacanth. Thus, the door has been closed on this important evolutionary question.

Molecular Clocks Measure the Rate of Evolutionary Change

In many cases, we would like to estimate not only which members of a set of species are most closely related, but also when their common ancestors lived. The ability to construct phylogenetic trees from protein and nucleic acid sequences led to the development of **molecular clocks,** which use the rate of change in amino acid or nucleotide sequences as a way to estimate the time of divergence from a common ancestor.

To be useful, molecular clocks must be carefully calibrated. Molecular clocks can only measure changes in amino acids or nucleotides; they are linear over certain time scales, and times and dates must be added to

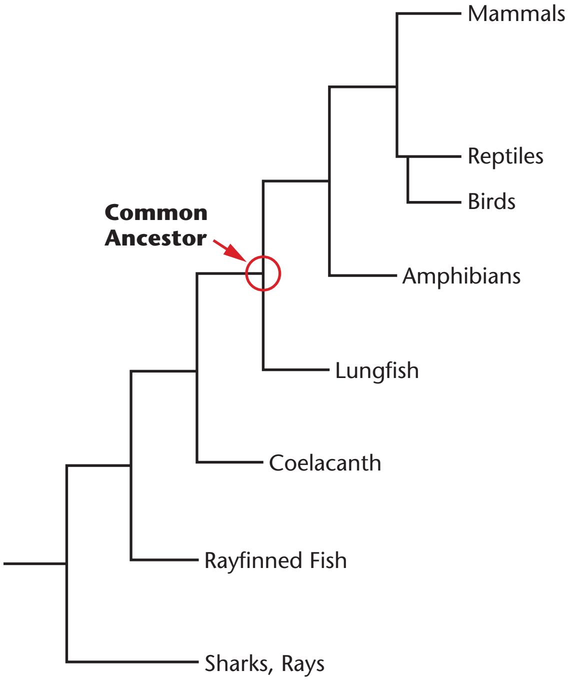


FIGURE 26.23 A phylogenetic tree of selected jawed vertebrates, including the lungfish and the coelacanth, shows that the lungfish shares the most recent common ancestor with these vertebrates.

the clock using independent evidence such as the fossil record. Figure 26.24 shows a molecular clock showing divergence times from a common ancestor for humans and other vertebrates based on the fossil record [Figure 26.24(a)] and molecular data [26.24(b)]. In both cases, changes in amino acid sequence and nucleotide sequence increase linearly with time. The results show that humans and zebrafish last shared a common ancestor about 450 million years ago, and humans last shared a common ancestor with chimpanzees about 7-10 million years ago.

The Complex Origins of the Human Genome

Current fossil, molecular, and genomic evidence indicates that our species, *Homo sapiens*, arose in Africa about 300,000 years ago from earlier species of *Homo*. When populations of *H. sapiens* first expanded out of Africa sometime between 50,000 and 70,000 years ago, parts of Europe and Asia were already occupied by members of other human (*Homo*) species. Advances in DNA-sequencing technology and new methods of DNA extraction that allow the recovery of genomic DNA from fossil remains have created a new field, called **paleogenomics,** which in turn, has revolutionized the study of human evolution. The genomes of two extinct groups who lived in the Middle East, Asia, and Europe, the Neanderthals and the Denisovans, have been sequenced and compared with the genomes of present-day humans. The results show that modern human populations outside Africa, including those of the Middle East, Europe, Asia, Australia/Oceania, and the Americas, carry sequences from these two groups. We know quite a lot about the genome of the Neanderthals and the contributions they made to our genome.

The first Neanderthal genome was assembled in 2010 from three skeletons discovered in a Croatian cave. Since then, genomes from several other Neanderthals have been sequenced. Comparative genome analysis shows that the genomes of our species and the Neanderthals are the same size (about 3.2 billion base pairs) and are 99.7 percent identical.

Populations of *H. neanderthalensis* lived in Europe and western Asia from some 300,000 years ago until they disappeared about 40,000 years ago. For at least 30,000 years, Neanderthals coexisted with anatomically modern humans (*H. sapiens*) in regions of the Middle East and Europe, providing an opportunity for interbreeding between these species. In fact, gene flow from extinct Neanderthals to modern humans through interbreeding is estimated to represent about 2 percent of the genome of non-African populations. Thus, the 99.7 percent sequence identity between the two species includes the 2 percent contributed by Neanderthals that has become fixed in the genome of our species. However, different individuals carry different portions of the Neanderthal genome; taken together, upward of 20 percent of the Neanderthal genome may be present in the genomes of modern non-African populations.

From these studies, two conclusions can be drawn. First, Neanderthals are not direct ancestors of our species. Second, Neanderthals and members of our species did interbreed, and Neanderthals contributed to our genome. Thus, although Neanderthals are extinct, some of their DNA has survived and is a fixed part of our genome. Based

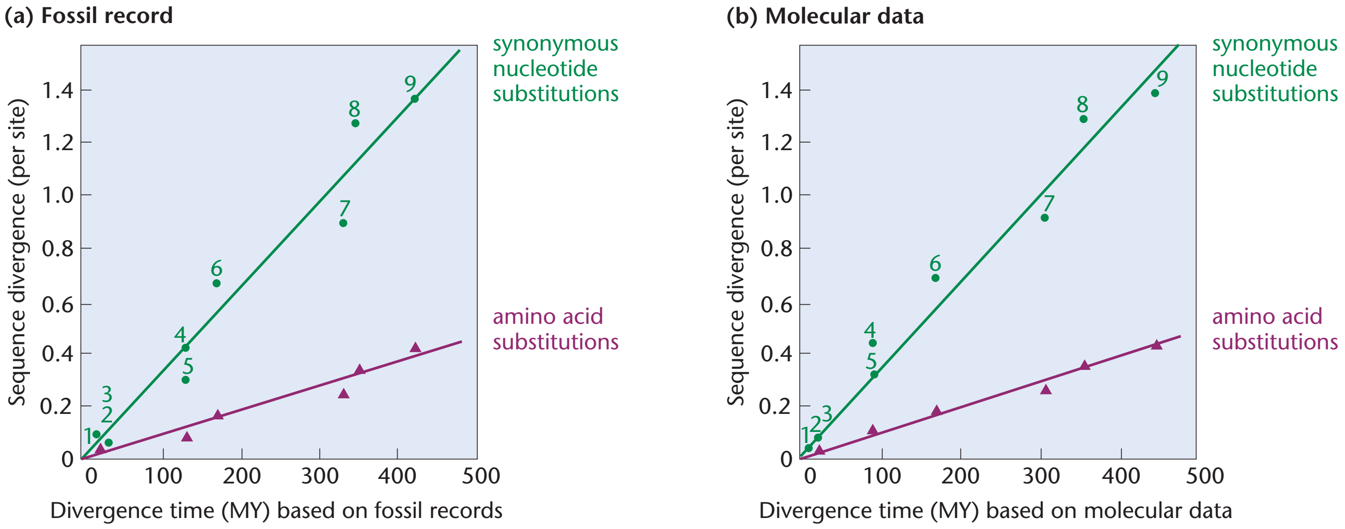


FIGURE 26.24 The relationship between the number of amino acid substitutions and the number of nucleotide substitutions for 4198 nuclear genes from 10 vertebrate species. Humans versus (1) chimpanzee, (2) orangutan, (3) macaque, (4) mouse, (5) cow, (6) opossum, (7) chicken, (8) western clawed frog, and (9) zebrafish. In (a) the data are calculated by divergence times based on the fossil record, and in (b) the data are based on synonymous nucleotide substitutions, which are mutations that do not result in any changes in the amino acid sequence of a protein. MY = millions of years.

on the size and distribution of Neanderthal DNA sequences in the genome of modern humans, it is estimated that mixing of these genomes occurred between 50,000 and 60,000 years ago.

In 2008, human fossils were discovered in a cave near Denisova, Siberia (**Figure 26.25**). A complete mtDNA genome sequence showed that these fossils belonged to a group separate from both Neanderthals and our species. They were named the Denisovans. A nuclear Denisovan genome sequence shows that they are more closely related to Neanderthals than to our species, and that Denisovans and Neanderthals separated from a common ancestral species more than 430,000 years ago. In addition, the Denisovan genome contains sequences from another, as yet unknown, archaic group that made no contribution to the Neanderthal genome.

Analysis of modern human populations shows that about 5 to 6 percent of the DNA in the genomes of Melanesian islanders in the South Pacific is derived from the Denisovans. Smaller amounts of Denisovan DNA are found in the genomes of Australian aborigines, as well as Polynesians, Fujians, east Indonesians, and some populations of East Asia. As things stand now, we know that as a result of gene flow, some members of our species outside of Africa carry DNA from one or two other human groups (**Figure 26.26**). Recent work has determined that Denisovan DNA sequences entered our genome about 44,000–54,000 years ago, roughly 6000 years after our genome mixed with that of Neanderthals.

The Neanderthal and Denisovan genomes were assembled from fossil remains that are 40,000 to 80,000 years old. The recent sequencing of a genome from a 700,000-year-old horse fossil opens the possibility that genome sequences can be recovered from fossils of much older human species and used to identify the archaic species that contributed to the Denisovan genome. For now, using the paleogenomic techniques currently available, we can expect exciting answers to questions about the similarities and differences between our genome and those of other human species, providing revolutionary insights into the evolution of our species and other human species that preceded us on this planet.

FIGURE 26.25 The cave in Denisova, Siberia, where the Denisovan fossils were discovered.

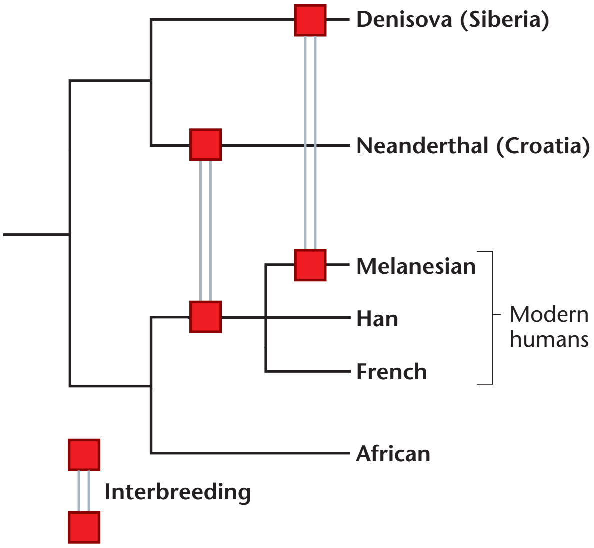


FIGURE 26.26 A phylogenetic tree showing the relationships among modern humans, Neanderthals, and Denisovans. The latter two groups branched off from our last common ancestor before our species left Africa. Genomic analysis shows that there was interbreeding between members of our species with Neanderthals and Denisovans, making our genome a mosaic with contributions from at least two other human species.

GENETICS, ETHICS, AND SOCIETY

Tracking Our Genetic Footprints out of Africa

Approximately 2 million years ago, a large-brained, tool-using hominid called *Homo erectus* appeared in East Africa. By 1.7 million years ago, *H. erectus* had spread into Eurasia and South Asia. Most scientists agree that *H. erectus* likely developed into *H. heidelbergensis*—a species that became the ancestor to our species (in Africa), Neanderthals (in Europe), and Denisovans (in Asia). These groups disappeared 50,000 to 30,000 years ago—around the same time that anatomically modern humans (*H. sapiens*) appeared all over the world.

At present, the most widely accepted hypothesis explaining the presence of anatomically modern humans is the out-of-Africa hypothesis. This hypothesis is based on genetic data derived from mitochondrial, Y chromosome, and whole-genome sequencing of both archaic hominin fossils and modern human populations. The out-of-Africa hypothesis states that *H. sapiens* evolved from the descendants of *H. heidelbergensis* in Africa about 300,000 years ago. Around 50,000 years ago, a small band of *H. sapiens* (perhaps fewer than 1000) left Africa. By 40,000 years ago, they had reached Europe, Asia, and Australia. In the out-of-Africa model, *H. sapiens* interbred with Neanderthal and Denisovan populations, and then became the only species in the genus by about 30,000 years ago.

Most genetic evidence appears to support the out-of-Africa hypothesis. Humans all over the globe are extremely similar genetically. DNA sequences from any two humans chosen at random are 99.9 percent identical. More genetic identity exists between two persons chosen at random from a human population than between two chimpanzees chosen at random from a chimpanzee population. Interestingly, about 90 percent of the genetic differences that do exist occur between individuals rather than between populations. This unusually high degree of genetic relatedness in all humans supports the idea that our species arose recently from a small founding group of individuals. Other genetic data show that the highest levels of human genetic variation occur within African populations. This implies that the earliest branches of *H. sapiens* diverged in Africa and had a longer time to accumulate DNA mutations.

As with any explanation of human origins, the out-of-Africa hypothesis is actively debated. Some data suggest two or more out-of-Africa dispersals, as well as different timings of dispersals and migration routes. As methods for sequencing DNA from ancient fossils improve, it will soon be possible to fill the gaps in our understanding of the genetic pathways leading out of Africa and to resolve age-old questions about our origins.

Your Turn

Take time, individually or in groups, to consider the following questions. Investigate the references and links dealing with the ethical and technological aspects of how we understand the origins of modern humans.

1. Some genetic and archaeological evidence appears to support two separate dispersals of humans out of Africa. What are these data, and how might they be reconciled with the single-dispersal hypothesis?

Start your investigations by reading Nielsen, R. et al. (2017). Tracing the peopling of the world through genomics. *Nature* 541:302–310, *and* Tucci, S. and Akey, J. M. (2016). A map of human wanderlust. *Nature* 538:179–180.

2. Given that genetic studies show that all people on Earth are remarkably similar genetically, how did we come to develop the concept of racial differences? How has modern genomics contributed to the debate about the validity and definition of the term “race”?

For an interesting discussion of race, human variation, and genomics, see Lewontin, R. C. (2006). Confusion about human races, *on the Social Sciences Research Center Web site* (raceandgenomics.ssrc.org/Lewontin). *Also, see* Cooper, R. S. (2013). Race in biological and biomedical research. *Cold Spring Harb. Perspect. Med.* 3(11): a008573.

CASE STUDY A Tale of Two Olivias

Olivia S. was born with a rare recessive disorder called tyrosinemia. The next day, Olivia M. was born in a neighboring state with the same disorder. Tyrosinemia is caused by the lack of an enzyme in the degradation pathway of the amino acid tyrosine. Accumulation of metabolic intermediates causes progressive liver dysfunction and kidney problems. One-year-old Olivia S. is healthy and has no symptoms of the disorder. At the same age, Olivia M. developed total liver failure. Olivia S. was born in a state where newborns are tested for tyrosinemia, but Olivia M. was born in a state where newborns are not tested for this disorder. A week after diagnosis, Olivia S. was placed on a low-tyrosine diet and prescribed a drug to block the accumulation of metabolic intermediates. Olivia M. was not diagnosed until she was in liver failure; she then was placed on a low-tyrosine diet, was prescribed medication, and underwent a liver transplant. She faces a lifetime of antirejection drug therapy and may require a kidney transplant. In the United States, newborn screening programs are developed independently by each state and are often based on a cost–benefit analysis to decide which diseases are included in testing. In the United States, tyrosinemia occurs in only 1/100,000 births, and in this case, two states made different decisions about newborn testing for this disorder.

1. In a region of Quebec, Canada, 1 in 22 people are heterozygous for the mutant tyrosinemia allele. Using the frequency of heterozygotes, calculate the frequency of recessive homozygotes in this population. What might explain the difference between the frequency of tyrosinemia in the U.S. population and in this particular Canadian population?
2. Critics argue that a uniform panel of disorders should be used by all states in newborn testing. Aside from cost–benefit ratios, what would you regard as ethical guidelines for use in deciding which disorders to include or exclude in a newborn testing program?
3. Others argue that the current testing system should be replaced by whole genome sequencing for all newborns. What do you see as the ethical pros and cons of this position?

See Tarini, B., and Goldenberg, A. (2012). Ethical issues with newborn screening in the genomics era. *Ann. Rev. Genomics Hum Genet.* 13:381–393.

Summary Points

Mastering Genetics For activities, animations, and review quizzes, go to the Study Area.

1. Genetic variation is a characteristic of most populations. In some cases, this can be observed at the phenotypic level, but analysis at the amino acid level and especially the nucleotide level provides a more direct way to estimate genetic variation.
2. Using the assumptions of the Hardy–Weinberg law, it is possible to estimate allele and genotype frequencies in populations.
3. The Hardy–Weinberg law can be used to determine allele and genotype frequencies for genes with multiple alleles and for genes on the X chromosome. In addition, this method can be used to calculate the frequency of heterozygotes for a given gene in a population.
4. Natural selection changes allele frequency in populations leading to evolutionary change. Selection for quantitative traits can involve directional selection, stabilizing selection, or disruptive selection.
5. In addition to natural selection, other forces act on allele frequencies in populations. These include mutation, migration, and genetic drift. Nonrandom mating alters genotype frequencies but does not change allele frequencies.
6. The formation of new species depends on the formation of subpopulations and the accumulation of enough genetic differences that, when reunited, members of the subpopulations cannot interbreed.
7. Phylogenetic analysis using morphology, amino acid sequences, or nucleotide sequences can be used to construct phylogenetic trees showing the evolutionary relationships among a group of organisms. When calibrated with molecular clocks, the evolutionary changes on a phylogenetic tree can be calibrated with a time scale.
8. Phylogenetic analysis combined with genome sequence data from humans, Neanderthals, and Denisovans has helped scientists reconstruct the complex origin of our species' genome.

INSIGHTS AND SOLUTIONS

1. Tay–Sachs disease is caused by loss-of-function mutations in a gene on chromosome 15 that encodes a lysosomal enzyme. Tay–Sachs is inherited as an autosomal recessive condition. Among Ashkenazi Jews of Central European ancestry, about 1 in 3600 children is born with the disease. What fraction of the individuals in this population are carriers?

Solution:
If we let p represent the frequency of the wild-type enzyme allele and q the total frequency of recessive loss-of-function alleles, and if we assume that the population is in Hardy–Weinberg equilibrium, then the frequencies of the genotypes are given by p^2 for homozygous normal, $2pq$ for carriers, and q^2 for individuals with Tay–Sachs. The frequency of Tay–Sachs alleles is thus

$$q = \sqrt{q^2} = \sqrt{\frac{1}{3600}} = 0.017$$

Since $p + q = 1$, we have

$$p = 1 - q = 1 - 0.017 = 0.983$$

Therefore, we can estimate that the frequency of carriers is

$$2pq = 2(0.983)(0.017) = 0.033$$

or about 1 in 30.

2. A single plant twice the size of others in the same population suddenly appears. Normally, plants of that species reproduce by self-fertilization and by cross-fertilization. Is this new giant plant simply a variant, or could it be a new species? How would you determine which it is?

Solution:
One of the most widespread mechanisms of speciation in higher plants is polyploidy, the multiplication of entire sets of chromosomes. The result of polyploidy is usually a larger plant with larger flowers and seeds. There are two ways of testing the new variant to determine whether it is a new species. First, the giant plant should be crossed with a normal-sized plant to see whether the giant plant produces viable, fertile offspring. If it does not, then the two different types of plants would appear to be reproductively isolated. Second, the giant plant should be cytogenetically screened to examine its chromosome complement. If it has twice the number of its normal-sized neighbors, it is a tetraploid that may have arisen spontaneously. If the chromosome number differs by a factor of two and the new plant is reproductively isolated from its normal-sized neighbors, it is a new species.

Problems and Discussion Questions

1. HOW DO WE KNOW? Population geneticists study changes in the nature and amount of genetic variation in populations, the distribution of different genotypes, and how forces such as selection and drift act on genetic variation to bring about evolutionary change in populations and the formation of new species. From the explanation given in the chapter, what answers would you propose to the following fundamental questions?

(a) How do we know how much genetic variation is in a population?

(b) How do geneticists detect the presence of genetic variation as different alleles in a population?

(c) How do we know whether the genetic structure of a population is static or dynamic?

(d) How do we know when populations have diverged to the point that they form two different species?

(e) How do we know the age of the last common ancestor shared by two species?

2. CONCEPT QUESTION Read the Chapter Concepts list on page 621. All these pertain to the principles of population genetics and the evolution of species. Write a short essay describing the roles of mutation, migration, and selection in bringing about speciation.

3. Price et al. [(1999). *J. Bacteriol.* 181:2358–2362] conducted a genetic study of the toxin transport protein (PA) of *Bacillus anthracis*, the bacterium that causes anthrax in humans. Within the 2294-nucleotide gene in 26 strains they identified five point mutations—two missense and three synonyms—among different isolates. Necropsy samples from an anthrax outbreak in 1979 revealed a novel missense mutation and five unique nucleotide changes among ten victims. The authors concluded that these data indicate little or no horizontal transfer between different *B. anthracis* strains.

(a) Which types of nucleotide changes (missense or synonyms) cause amino acid changes?

(b) What is meant by "horizontal transfer"?

(c) On what basis did the authors conclude that evidence of horizontal transfer is absent from their data?

4. The genetic difference between two *Drosophila* species, *D. heteroneura* and *D. silvestris,* as measured by nucleotide diversity, is about 1.8 percent. The difference between chimpanzees (*Pan troglodytes*) and humans (*H. sapiens*) is about the same, yet the latter species is classified in a different genera. In your opinion, is this valid? Explain why.

5. The use of nucleotide sequence data to measure genetic variability is complicated by the fact that the genes of many eukaryotes are complex in organization and contain 5′ and 3′ flanking regions as well as introns. Researchers have compared the nucleotide sequence of two cloned alleles of the *γ-globin* gene from a single individual and found a variation of 1 percent. Those differences include 13 substitutions of one nucleotide for another and three short DNA segments that have been inserted in one allele or deleted in the other. None of the changes takes place in the gene's exons (coding regions). Why do you think this is so, and should it change our concept of genetic variation?

6. Consider rare disorders in a population caused by an autosomal recessive mutation. From the frequencies of the disorder in the population given, calculate the percentage of heterozygous carriers:

(a) 0.0064

(b) 0.000081

(c) 0.09

(d) 0.01

(e) 0.10

7. What must be assumed in order to validate the answers in Problem 7?

8. In a population where only the total number of individuals with the dominant phenotype is known, how can you calculate the percentage of carriers and homozygous recessives?

9. If 4 percent of a population in equilibrium expresses a recessive trait, what is the probability that the offspring of two individuals who do not express the trait will express it?

10. Consider a population in which the frequency of allele *A* is $p = 0.7$ and the frequency of allele *a* is $q = 0.3$, and where the alleles are codominant. What will be the allele frequencies after one generation if the following occurs?

(a) $w_{AA} = 1, w_{Aa} = 0.9, w_{aa} = 0.8$

(b) $w_{AA} = 1, w_{Aa} = 0.95, w_{aa} = 0.9$

(c) $w_{AA} = 1, w_{Aa} = 0.99, w_{aa} = 0.98$

(d) $w_{AA} = 0.8, w_{Aa} = 1, w_{aa} = 0.8$

11. If the initial allele frequencies are $p = 0.5$ and $q = 0.5$ and allele *a* is a lethal recessive, what will be the frequencies after 1, 5, 10, 25, 100, and 1000 generations?

12. Under what circumstances might a lethal dominant allele persist in a population?

13. Assume that a recessive autosomal disorder occurs in 1 of 10,000 individuals (0.0001) in the general population and that in this population about 2 percent (0.02) of the individuals are carriers for the disorder. Estimate the probability of this disorder occurring in the offspring of a marriage between first cousins. Compare this probability to the population at large.

14. One of the first Mendelian traits identified in humans was a dominant condition known as *brachydactyly*. This gene causes an abnormal shortening of the fingers or toes (or both). At the time, some researchers thought that the dominant trait would spread until 75 percent of the population would be affected (because the phenotypic ratio of dominant to recessive is 3 : 1). Show that the reasoning was incorrect.

15. Describe how populations with substantial genetic differences can form. What is the role of natural selection?

16. Achondroplasia is a dominant trait that causes a characteristic form of dwarfism. In a survey of 50,000 births, five infants with achondroplasia were identified. Three of the affected infants had affected parents, while two had normal parents. Calculate the mutation rate for achondroplasia and express the rate as the number of mutant genes per given number of gametes.

17. A recent study examining the mutation rates of 5669 mammalian genes (17,208 sequences) indicates that, contrary to popular belief, mutation rates among lineages with vastly different generation lengths and physiological attributes are remarkably constant [Kumar, S., and Subramanian, S. (2002). *Proc. Natl. Acad. Sci. USA* 99:803–808]. The average rate is estimated at 12.2×10^{-9} per bp per year. What is the significance of this finding in terms of mammalian evolution?

18. What are considered significant factors in maintaining the surprisingly high levels of genetic variation in natural populations?

19. A botanist studying water lilies in an isolated pond observed three leaf shapes in the population: round, arrowhead, and scalloped. Marker analysis of DNA from 125 individuals showed the round-leaf plants to be homozygous for allele *r1*, while the plants with arrowhead leaves were homozygous for a different allele at the same locus, *r2*. Plants with scalloped leaves showed DNA profiles with both the *r1* and *r2* alleles. Frequency of the *r1* allele was estimated at 0.81. If the botanist counted 20 plants with scalloped leaves in the pond, what is the inbreeding coefficient *F* for this population?

20. A farmer plants transgenic Bt corn that is genetically modified to produce its own insecticide. Of the corn borer larvae feeding on these Bt crop plants, only 10 percent survive unless they have at least one copy of the dominant resistance allele *B* that confers resistance to the Bt insecticide. When the farmer first plants Bt corn, the frequency of the *B* resistance allele in the corn borer population is 0.02. What will be the frequency of the resistance allele after one generation of corn borers have fed on Bt corn?

21. In an isolated population of 50 desert bighorn sheep, a mutant recessive allele *c* when homozygous causes curled coats in both males and females. The normal dominant allele *C* produces straight coats. A biologist studying these sheep counts four with curled coats. She also takes blood samples from the population for DNA analysis, which reveals that 17 of the sheep are heterozygous carriers of the *c* allele. What is the inbreeding coefficient *F* for this population?

22. To increase genetic diversity in the bighorn sheep population described in Problem 23, ten sheep are introduced from a population where the *c* allele is absent. Assuming that random mating occurs between the original and the introduced sheep, and that the *c* allele is selectively neutral, what will be the frequency of *c* in the next generation?

23. What genetic changes take place during speciation?

24. Some critics have warned that the use of gene therapy to correct genetic disorders will affect the course of human evolution. Evaluate this criticism in light of what you know about population genetics and evolution, distinguishing between somatic gene therapy and germ-line gene therapy.

25. List the barriers that prevent interbreeding, and give an example of each.

26. What are the two groups of reproductive isolating mechanisms? Which of these is regarded as more efficient, and why?

Extra-Spicy Problems

27. A form of dwarfism known as Ellis–van Creveld syndrome was first discovered in the late 1930s, when Richard Ellis and Simon van Creveld shared a train compartment on the way to a pediatrics meeting. In the course of conversation, they discovered that they each had a patient with this syndrome. They published a description of the syndrome in 1940. Affected individuals have a short-limbed form of dwarfism and often have defects of the lips and teeth, and polydactyly (extra fingers). The largest pedigree for the condition was reported in an Old Order Amish population in eastern Pennsylvania by Victor McKusick and his colleagues (1964). In that community, about 5 per 1000 births are affected, and in the population of 8000, the observed frequency is 2 per 1000. All affected individuals have unaffected parents, and all affected cases can trace their ancestry to Samuel King and his wife, who arrived in the area in 1774. It is known that neither King nor his wife was affected with the disorder. There are no cases of the disorder in other Amish communities, such as those in Ohio or Indiana.

(a) From the information provided, derive the most likely mode of inheritance of this disorder. Using the Hardy–Weinberg law, calculate the frequency of the mutant allele in the population and the frequency of heterozygotes, assuming Hardy–Weinberg conditions.

(b) What is the most likely explanation for the high frequency of the disorder in the Pennsylvania Amish community and its absence in other Amish communities?

28. The original source of new alleles, upon which selection operates, is mutation, a random event that occurs without regard to selectional value in the organism. Although many model organisms have been used to study mutational events in populations, some investigators have developed abiotic molecular models. Soll et al. (2006. *Genetics* 175: 267-275) examined one such model to study the relationship between both deleterious and advantageous mutations and population size in a ligase molecule composed of RNA (a ribozyme). Soll found that the smaller the population of molecules, the more likely it was that not only deleterious mutations but also advantageous mutations would disappear. Why would population size influence the survival of both types of mutations (deleterious and advantageous) in populations?

29. A number of comparisons of nucleotide sequences among hominids and rodents indicate that inbreeding may have occurred more often in hominid than in rodent ancestry. Bakewell et al. (2007. *Proc. Nat. Acad. Sci. [USA]* 104: 7489-7494) suggest that an ancient population bottleneck that left approximately 10,000 humans might have caused early humans to have a greater chance of genetic disease. Why would a population bottleneck influence the frequency of genetic disease?

30. Shown below are two homologous lengths of the alpha and beta chains of human hemoglobin. Consult a genetic code dictionary (Figure 13.7), and determine how many amino acid substitutions may have occurred as a result of a single nucleotide substitution. For any that cannot occur as a result of a single change, determine the minimal mutational distance.

Alpha: ala val ala his val asp asp met pro
Beta: gly leu ala his leu asp asn leu lys

31. Recent reconstructions of evolutionary history are often dependent on assigning divergence in terms of changes in amino acid or nucleotide sequences. For example, a comparison of cytochrome c shows 10 amino acid differences between humans and dogs, 24 differences between humans and moths, and 38 differences between humans and yeast. Such data provide no information as to the absolute times of divergence for humans, dogs, moths, and yeast. How might one calibrate the molecular clock to an absolute time clock? What problems might one encounter in such a calibration?

SPECIAL TOPICS IN MODERN GENETICS 1

CRISPR-Cas and Genome Editing

Genetic research is often a slow incremental process that may extend our understanding of a concept or improve the efficiency of a genetic technology. More rarely, discoveries advance the field in sudden and profound ways. For example, studies in the early 1980s led to the discovery of catalytic RNAs, which transformed how geneticists think about RNA. Around the same time, the development of the polymerase chain reaction (PCR) provided a revolutionary tool for geneticists and other scientists. Rapid and targeted DNA amplification is now indispensable to genetic research and medical science. Given this context, one can appreciate how rare and significant a discovery would be that both illuminates a novel genetic concept as well as yields a new technology for genetics research and application. CRISPR-Cas is exactly that.

For over a century, scientists have studied the biological warfare between bacteria and the viruses that infect them. However, in 2007, experiments confirmed that bacteria have a completely novel defense mechanism against viruses known as CRISPR-Cas. This discovery completely changed the scope of our understanding of how bacteria and viruses combat one another, and coevolve. Moreover, the CRISPR-Cas system has now been adapted as an incredibly powerful tool for genome editing.

The ability to specifically and efficiently edit a genome has broad implications for research, biotechnology, and medicine. For decades, geneticists have used various strategies for genome editing with many successes, but also with limited efficiency and a significant investment of time and resources. CRISPR-Cas has been developed into an efficient, cost-effective molecular tool that can introduce precise and specific edits to a genome. It is not without its limitations, but it represents a technological leap, which we have not seen, arguably, since the innovation of PCR.

> "CRISPR-Cas has been developed into an efficient, cost-effective molecular tool that can introduce precise and specific edits to a genome."

The discovery of CRISPR-Cas has impacted genetics and other related fields at an unprecedented pace (**Figure ST 1.1**). CRISPR-Cas is the focus of numerous patent applications and disputes, has been approved for use in clinical trials to treat disease, has been used to edit the genome of human embryos as a proof of concept for future medical applications, has instigated international discussions on its ethical use, and is most deserving of its own chapter in a genetics textbook.

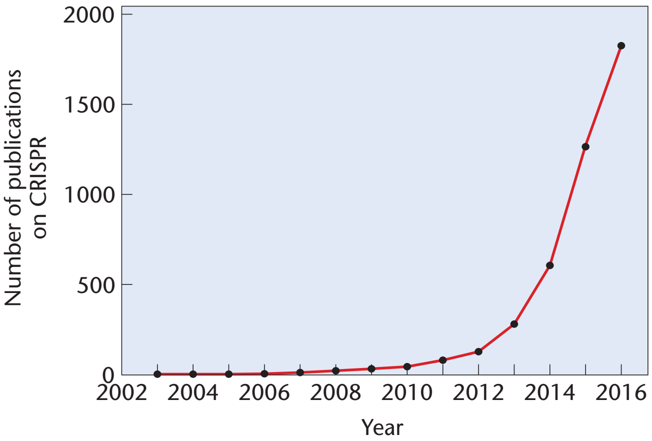


FIGURE ST 1.1 The number of publications returned in a search for "CRISPR" in PubMed by year.

ST 1.1 CRISPR-Cas Is an Adaptive Immune System in Prokaryotes

Bacteria and viruses (bacteriophages or phages) engage in constant biological warfare. Consequently, bacteria exhibit a diverse suite of defense mechanisms. For example, bacteria express endonucleases (restriction enzymes), which cleave specific DNA sequences. Such restriction enzymes destroy foreign bacteriophage DNA, while the bacterium protects its own DNA by methylating it. As you know (from Chapter 20), restriction enzymes have been adopted by molecular biologists for use in recombinant DNA technology. Bacteria can also defend against phage attack by blocking phage adsorption, blocking phage DNA insertion, and inducing suicide in infected cells to prevent the spread of infection to other cells. All of these defense mechanisms are considered **innate immunity** because they are not tailored to a specific pathogen.

In contrast, **adaptive immunity** refers to an evolving defense mechanism whereby past exposure to a pathogen stimulates improved defense against future exposure to the same pathogen. For example, when the human immune system is presented with a vaccine containing inactive virus, it “learns” how to defend against that same virus. It was thought that bacteria were too simple to possess such immunity. However, the discovery that CRISPR-Cas is an adaptive immune system in bacteria has overhauled our understanding of bacteria—phage warfare and their coevolution.

Discovery of CRISPR

A **CRISPR** is a genomic locus in bacteria that contains clustered regularly interspaced short palindromic repeats. Prior to the coining of this term, CRISPR loci were first identified in 1987 in the *Escherichia coli* genome based on a simple description of repeated DNA sequences with nonrepetitive *spacer* sequences between them. Since then, CRISPR loci have been identified in ~40 percent of bacteria species and in ~90 percent of archaea, another type of prokaryote (**Figure ST 1.2**). The spacers remained a mystery until 2005 when three independent studies demonstrated that CRISPR spacer sequences were identical to fragments of bacteriophage genomes. This insight led to speculation that viral sequences within CRISPR loci serve as a “molecular memory” of previous viral attacks.

The first experimental evidence that CRISPRs are important for adaptive immunity came from an unexpected place. Danisco, a Danish food science company, sought to create a strain of *Streptococcus thermophilus* that was more resistant to phage, thus making it more efficient for use in the production of yogurt and cheese. Philippe Horvath’s lab at Danisco, in collaboration with others, found that when they exposed *S. thermophilus* to a specific phage, bacterial cells that survived became resistant to the same phage strain, but not to other phage strains.[1] Furthermore, the resistant bacteria possessed new spacers within their CRISPR loci with an exact sequence match to portions of the genome of the phages by which they had been challenged.

Next, the Horvath lab showed that deletion of new spacers in the resistant strains abolished their phage resistance. Remarkably, the converse was also true; experimental insertion of new viral sequence-derived spacers into the CRISPR loci of sensitive bacteria rendered them resistant!

Finally, the Horvath lab noted that when bacteria were challenged by the same phage they had recently acquired immunity to, some bacterial cells would succumb to phage infection. Sequencing of the phage genome from these samples revealed that it had mutated, resulting in single-nucleotide mismatches between the phage genome and the spacers in the bacterial genome that previously provided resistance. This cycle of bacteria evolving better defense mechanisms and phage evolving ways to evade them is an exquisite example of the biological arms race.

CRISPR-Cas is not limited to viral defense. Work from other labs showed that CRISPR-Cas may also target plasmids—extrachromosomal autonomous DNA molecules (see Chapter 6).

The CRISPR-Cas Mechanism for RNA-Guided Destruction of Invading DNA

Studies from several labs have now elucidated the mechanism underlying the bacterial adaptive immune system. In addition to the CRISPR loci, adaptive immunity is also dependent on a set of adjacent **CRISPR-associated (*cas*) genes**. The *cas*

[1] Key papers referenced in this Special Topic are listed in the end-of-chapter Selected Readings and Resources.

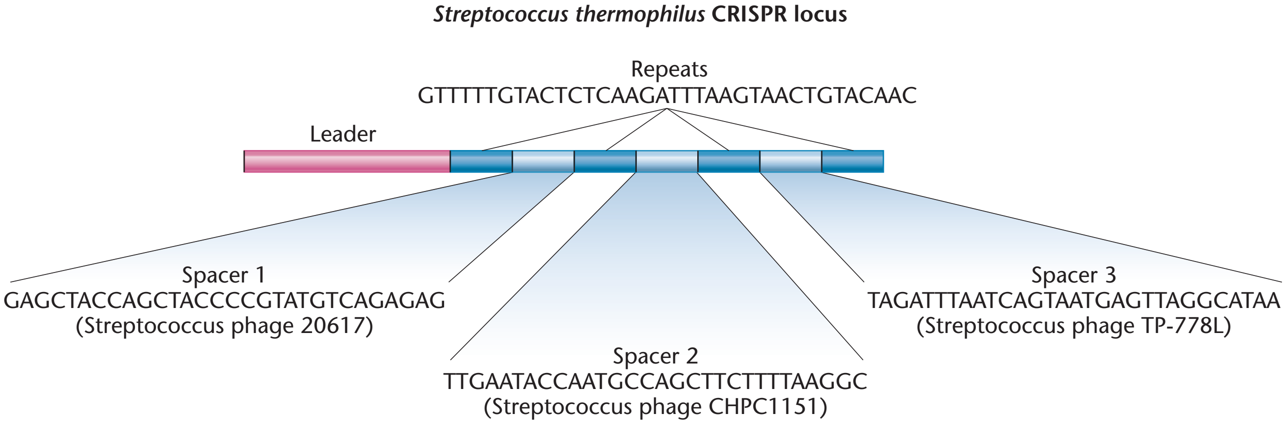

Streptococcus thermophilus CRISPR locus
Repeats
GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Leader
Spacer 1
GAGCTACCAGCTACCCCGTATGTCAGAGAG
(Streptococcus phage 20617)
Spacer 2
TTGAATACCAATGCCAGCTTCTTTTAAGGC
(Streptococcus phage CHPC1151)
Spacer 3
TAGATTTAATCAGTAATGAGTTAGGCATAA
(Streptococcus phage TP-778L)

FIGURE ST 1.2 A CRISPR locus from the bacterium *Streptococcus thermophilus* (LMG18311). Spacer sequences are derived from portions of bacteriophage genomes and are flanked on either side by a repeat sequence. Only 3 of 33 total spacers in this CRISPR locus are shown.

genes encode a wide variety of Cas proteins such as DNases, RNases, and proteins of unknown function. The **CRISPR-Cas** mechanism includes three steps outlined in **Figure ST 1.3**.

1. The first step is known as **spacer acquisition**. Invading phage DNA is cleaved into smaller fragments known as *protospacers*, which are then inserted into CRISPR loci to become new spacers. The Cas1 nuclease and an associated Cas2 protein control this process. New spacers are inserted proximal to the *leader sequence* of the CRISPR locus, with older spacers being located progressively more distal. When new spacers are added to the CRISPR locus, repeat sequences are duplicated such that each spacer is flanked by repeats on each side.
2. In the second step, CRISPR loci are transcribed, starting at the promoter within the leader, into long transcripts. These transcripts are then processed into short **CRISPR-derived RNAs (crRNAs)**, each containing a single spacer flanked on both sides by repeat sequences. This step is referred to as **crRNA biogenesis**. The mechanistic details and the Cas proteins and RNases required for this step vary (see below).
3. The third and final step is referred to as **target interference**. Mature crRNAs associate with Cas nucleases, or nuclease complexes, and recruit them to complementary sequences in invading phage DNA. The Cas nucleases then cleave the viral DNA, thus neutralizing infection.

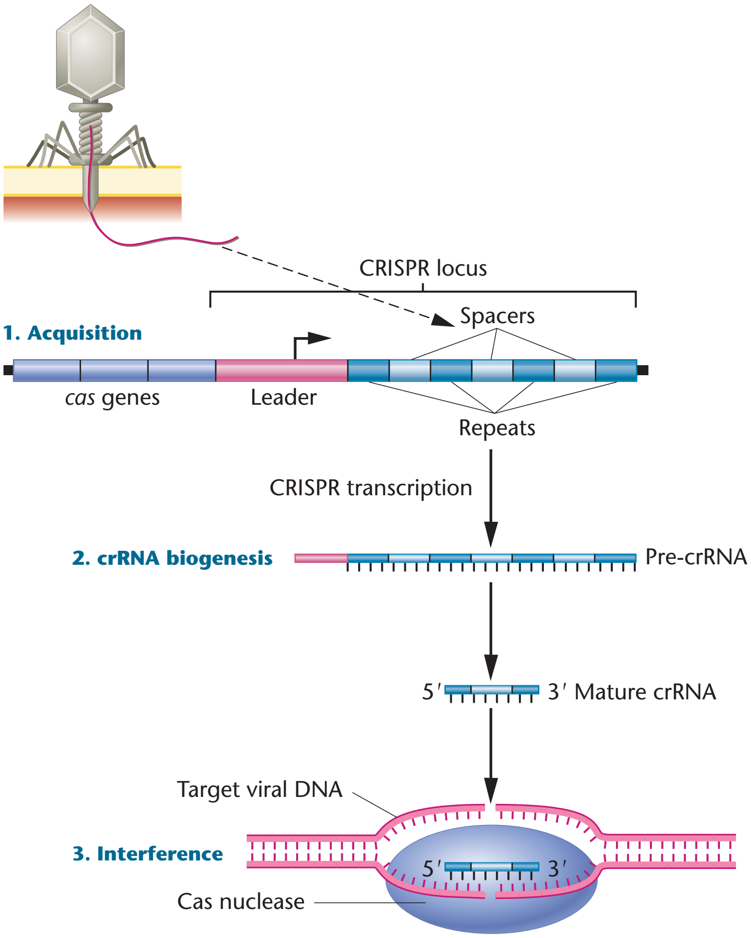


FIGURE ST 1.3 The mechanism of CRISPR-Cas adaptive immunity in bacteria. CRISPR loci contain spacer sequences derived from viral DNA separated by repeats, and a leader sequence. Nearby *cas* genes encode proteins that are also involved. The CRISPR-Cas mechanism has three steps: (1) *spacer acquisition* (viral DNA is inserted into CRISPR loci), (2) *crRNA biogenesis* (CRISPR loci are transcribed and processed into short crRNAs), and (3) *target interference* (viral DNA is cleaved by crRNA-guided Cas proteins).

Type II CRISPR-Cas Systems

Several different types of CRISPR-Cas systems (types I, II, III, etc.) are used by different species of bacteria or archaea. They all employ a crRNA-guided nuclease, or nuclease complex, which destroys complementary viral DNA. However, different Cas proteins and mechanistic details define each type. For example, *E. coli* has a type I system characterized by crRNAs with stem-loops and a nuclease complex called Cascade. *Staphylococcus aureus*, a bacterium that causes skin infections, has a type III system characterized by a Cas10-containing nuclease complex. However, of all systems, the type II CRISPR-Cas system of *Streptococcus pyogenes*, the bacterium that causes strep throat, is the best studied due to its simplicity.

The CRISPR-Cas system in *S. pyogenes* requires only a few essential components. Central to its simplicity is the **Cas9** nuclease. Whereas type I and II systems require multi-subunit protein complexes to mediate RNA-guided viral DNA destruction during the interference step, the single Cas9 protein is sufficient in *S. pyogenes*. Furthermore, Cas9 also plays a role in spacer acquisition and crRNA biogenesis.

The role of Cas9 during spacer acquisition is to cleave invading viral DNA. This process is not random; Cas9 selects protospacer sequences flanked by a **protospacer adjacent motif (PAM)** defined by the sequence 5′-NGG-3′, where N is any nucleotide. The selection of protospacers next to, but not including, a PAM sequence is important for the interference step. After protospacer cleavage, the Cas1/Cas2 complex integrates the DNA into a CRISPR locus.

Next, processing of the long pre-crRNA transcript into mature crRNAs requires the help of a noncoding RNA called a **transactivating crRNA (tracrRNA)**. The tracrRNA, which is transcribed from a different locus, binds to crRNA repeat regions through complementary base pairing. This complex is then bound by Cas9 and is cleaved into mature crRNA/tracrRNA duplexes by an RNase that specifically recognizes double-stranded RNA (RNase III).

The roles of Cas9, the crRNA, and the tracrRNA during interference are depicted in **Figure ST 1.4(a)**. The mature crRNA is bound to the tracrRNA by repeat region complementarity, while the single-stranded spacer region is used to recruit Cas9 to a complementary target sequence in the viral genome. Cas9 has two nuclease domains that cleave the target DNA sequence. The **HNH domain** cleaves the strand of viral DNA that is complementary to the crRNA, while the **RuvC domain** cleaves the noncomplementary strand. However, Cas9 will only cleave the DNA if the target is adjacent to a PAM sequence. Cas9's ability to discriminate between targets with or without a PAM provides this system with a way to distinguish "self" DNA from foreign DNA. Although the CRISPR spacers in the bacterial genome have perfect complementarity to crRNAs that are transcribed from them, CRISPR loci are not cut because they lack a PAM sequence.

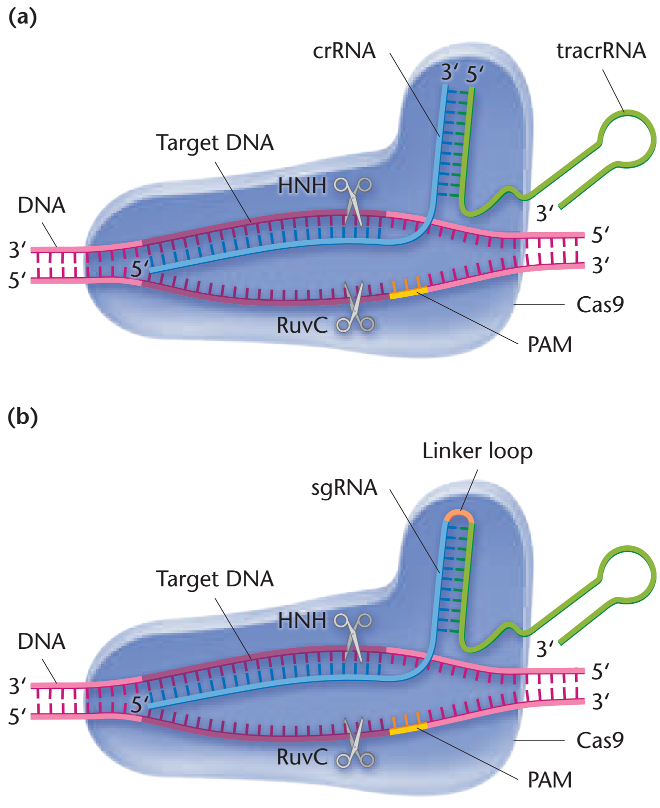


FIGURE ST 1.4 Cas9-mediated cleavage of target DNA. (a) Cas9 is guided by the crRNA/tracrRNA duplex to target sequences in the viral genome based on complementarity to the crRNA and the presence of a PAM sequence on the noncomplementary target DNA strand. The HNH domain of Cas9 cuts the complementary strand, while the RuvC domain cuts the other strand. (b) The CRISPR-Cas9 system has been simplified for use as a genome-editing tool by linking together a target-specific crRNA and a portion of the tracrRNA into a single guide RNA (sgRNA).

ST 1.2 CRISPR-Cas has been Adapted as a Powerful Tool for Genome Editing

While the mechanism of CRISPR-Cas adaptive immunity was under investigation, several scientists wondered if this system could be harnessed for other purposes. In 2008, Luciano Marraffini and Erik Sontheimer were the first to speculate that the CRISPR-Cas system could be used to target specific DNA sequences for destruction outside of bacteria and archaea. The first steps toward this goal were *in vitro* studies of CRISPR-Cas using a type II system due to its simplicity.

CRISPR-Cas9 *In Vitro*

In 2012, two independent research teams demonstrated that specific DNA sequences could be targeted efficiently, *in vitro*, by designing crRNAs with a complementary sequence. Both groups showed that this system required only purified Cas9, tracrRNA, a custom-designed crRNA, and a target sequence to cleave. In addition, the research team led by Jennifer Doudna and Emmanuelle Charpentier had a brilliant idea to further simplify the type II Cas9 system *in vitro*.

Using recombinant DNA technology (Chapter 20), the team engineered a hybrid RNA molecule called a **single guide RNA (sgRNA)**, which joined a 20-nucleotide-long targeting sequence of a crRNA to minimal sequences of the tracRNA necessary for Cas9 function [see **Figure ST 1.4(b)**]. Essentially any sequence, with a PAM, could be targeted for cleavage *in vitro* with just two simple components: an sgRNA containing customized crRNA and tracrRNA-derived sequences, and Cas9. This simple system became the basis for CRISPR-Cas genome editing.

CRISPR-Cas9 Genome Editing of Mammalian Cells

Genome editing, based on the ability to target and manipulate specific DNA sequences with specific nucleases, is not a new science. The first knockout mouse (Chapter 20), in which a specific gene was targeted for inactivation, was created in 1989. Since then, engineered endonucleases, such as zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), improved the efficiency of genome editing. However, these techniques are often slow and expensive because it is difficult to engineer a nuclease—a protein—to bind and cut a specific DNA sequence. In contrast, the specificity of Cas9 is derived from complementary base pairing between the sgRNA and the target sequence. No complex protein engineering is needed; the 20-nucleotide sequence of the sgRNA complementary to a target sequence provides the specificity for Cas9. The only restriction is that the target must have an adjacent PAM (5′-NGG-3′) sequence, which occurs with a frequency of once every 16 base pairs in a random sequence.

The first use of CRISPR-Cas9 as a genome-editing tool came in 2013 when the laboratories of Feng Zhang and George Church each independently edited the genomes of cultured mammalian cells. Their general strategies were the same as shown in **Figure ST 1.5**. Both labs engineered plasmid expression vectors (Chapter 20) carrying genes encoding Cas9 and an sgRNA with a specific DNA targeting sequence and introduced them into mammalian cells. A few modifications to the *in vitro* system pioneered by the Doudna and Charpentier labs were needed. For example, since the native context of Cas9 is the cytoplasm of a bacterial cell, nuclear localization signals (NLSs) were added to Cas9 to ensure it could access the nucleus of mammalian cells. In addition, Zhang and Church found that the short tracrRNA sequences of sgRNAs that worked *in vitro* were inefficient in mammalian cells; thus longer tracrRNA sequences were needed.

The CRISPR-Cas9 system is capable of cutting the genome at a precise location, but how can this be used for genome editing—the engineering of specific substitutions, deletions, or additions to a region of the genome for research, biotechnological, or medical applications? The answer to this question requires a basic understanding of the endogenous double-strand break repair pathways in the eukaryotic cell (introduced in Chapter 15). Double-stranded breaks in the genome may be repaired by **nonhomologous end-joining (NHEJ)** or by **homology-directed repair (HDR)**. These are competing repair pathways; either may be used to repair any given incidence of double-stranded break. NHEJ simply involves the ligation of broken DNA fragments. This process is error prone and often results in small insertions or deletions (*indels*) at the repair site. HDR is less error prone and uses an undamaged homologous chromosome or sister chromatid as a template to correctly repair a broken chromosome.

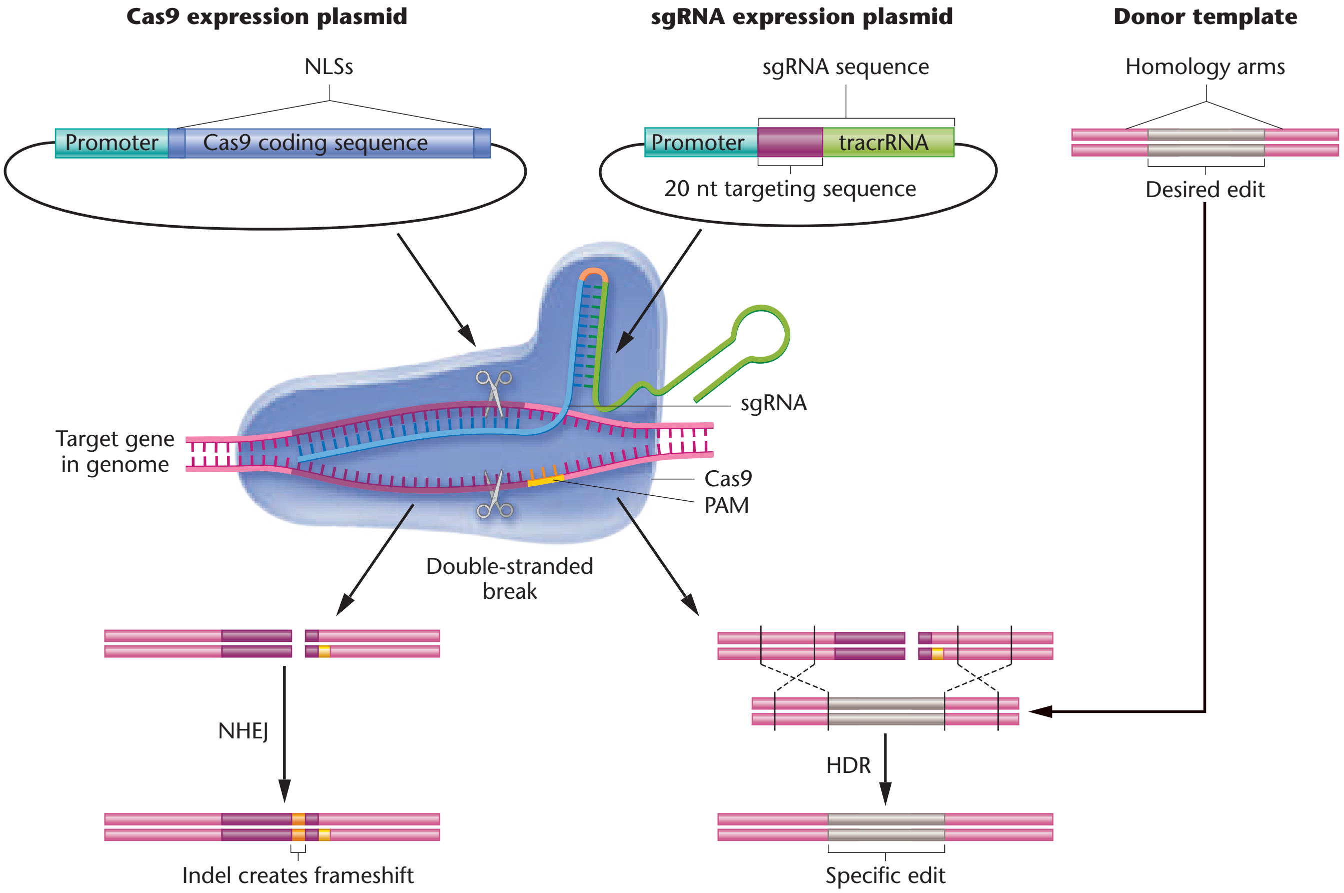


FIGURE ST 1.5 CRISPR-Cas9 genome editing in cultured mammalian cells. Cas9, with nuclear localization signals (NLSs), and an sgRNA are expressed from plasmids in mammalian cells. The sgRNA guides Cas9 to cleave a target site adjacent to a PAM sequence. The double-stranded DNA break can be repaired by NHEJ, which introduces insertions or deletions (indels), or by HDR. The latter mechanism can make specific edits using an introduced donor template.

If the goal of CRISPR-Cas9 genome editing is to disrupt a gene and create a nonfunctional allele, then simply adding Cas9 and an sgRNA into the cell may be sufficient. While the HDR mechanism may repair Cas9-induced double-stranded breaks correctly in some cells, in other cells the error-prone NHEJ pathway may introduce indels that result in a shift of the coding sequence reading frame, and thus lead to gene disruption. However, if the goal of CRISPR-Cas9 gene editing is to make a more precise edit, HDR can be "tricked" into using an artificial **donor template** (instead of the homologous chromosome) to make complex substitutions, deletions, or additions. The donor template is an experimentally introduced DNA molecule carrying a sequence with desired edits flanked by "homology arms" with sequences that match regions of the genome adjacent to the genomic target. Through the HDR mechanism, the target sequence in the genome is replaced by the sequence on the donor template. Examples of genome editing for research, biotechnology, and medicine are discussed later in this chapter.

By introducing Cas9, an sgRNA, and a donor template, the Zhang and Church labs both showed specific editing of several different targets in mammalian cells. For some targets, the Church lab reported successful editing in 25 percent of cells. Furtherore, in some cases, editing was detected only 20 minutes after adding the editing components!

Other innovations that improved CRISPR-Cas9 genome editing also came out of the Zhang and Church studies. As mentioned above, NHEJ often leads to frameshifts and thus inactivation of a gene; specific edits require HDR and a donor template. To increase the chances of genome editing by the HDR mechanism, the researchers took advantage of the fact that the eukaryotic cell uses the HDR pathway not just for double-stranded breaks in DNA, but also for single-stranded breaks. Therefore, the researchers engineered a mutation into the RuvC domain of Cas9. With only one functional nuclease domain, the HNH domain, Cas9 becomes a "nickase"—it cleaves only one of the two strands of the target DNA sequence. Cas9 nickase activity decreased the proportion of NHEJ edits and increased the proportion of HDR specific edits. Zhang and Church also showed that introduction of two different sgRNAs targeting different regions of the same gene on opposite strands could increase the chance of making a deletion in a gene.

CRISPR-Cas Infidelity

CRISPR-Cas is clearly a powerful tool with immense potential, but it does have limitations. In some cases, Cas9 not only cuts at the intended target but also at off-target sites in the genome. Off-target edits may be due to an sgRNA having more than one perfect match in the genome. So, CRISPR-Cas9 infidelity can be partly addressed through careful design of sgRNAs for a given target. Even with careful design, off-target edits do still happen. Why?

To answer this question it is important to recall that Cas9 evolved as a bacterial defense mechanism against viral attack. In this context, a modest degree of Cas9 infidelity is likely an *advantage* to the bacterial cell because it will enable the cell to defend itself against rapidly evolving viruses. A bacterium that acquires immunity to a virus will be resistant to a mutated strain of the virus if Cas9 can cut viral DNA with an imperfect match to a crRNA. Therefore, Cas9's off-target edits in the mammalian genome may simply be a consequence of Cas9's evolutionary habits. Furthermore, in its native context, Cas9 searches for a target sequence against a tiny viral genome and the 1.8-million-base-pair bacterial (*Streptococcus pyogenes*) genome. When Cas9 is used as a tool for editing a large mammalian genome, such as that of humans, it must search for a target within a genome that is almost 1800 times larger!

To address this problem, several labs have tried altering amino acids in Cas9 to improve its specificity. Others have designed Web-based algorithms to improve sgRNA design. Others still have turned back to bacteria and archaea looking for CRISPR systems with alternative enzymes to Cas9 that have improved fidelity or other desirable traits. Some studies have shown that the Cpf1 nuclease, from the CRISPR system of bacteria in the *Francisella* and *Prevotella* genera, exhibits lower off-target editing than Cas9. Improving the specificity of CRISP-Cas edits to the human genome will be important for the safety of medical applications of this technology.

ST 1.3 CRISPR-Cas Technology Has Diverse Applications

Shortly after the seminal studies by Doudna and Charpentier, Zhang, and Church, many labs reported success with genome editing in a wide range of model systems such as mice, *Drosophila*, *Caenorhabditis elegans*, zebrafish, and plants. The CRISPR-Cas genome-editing revolution is now well under way with myriad applications in research, biotechnology, and medicine. With so many applications, the patents on CRISPR-Cas technology are likely to be far-reaching and, occasionally, an extremely lucrative venture. See Box 1 for a discussion of the battle for CRISPR-Cas patents.

CRISPR-Cas as a Tool for Basic Genetic Research

CRISPR-Cas is already an indispensable tool for basic genetics research. A fundamental objective that geneticists often pursue is to determine the function of a gene that has

BOX 1

The CRISPR-Cas9 Patent Battle

A patent is a legal license to an invention that excludes others from using the invention without permission for 20 years. There are limits to what can be patented. According to the U.S. Patent and Trademark Office (USPTO), an invention is only patentable if it is novel, useful, and not obvious. Furthermore, it must not be "naturally occurring."

The CRISPR genomic feature and the Cas9 protein are naturally occurring in bacteria and thus cannot be patented directly. However, the various innovations that enable CRISPR-Cas9 to be a powerful tool for DNA manipulation are patentable. With myriad research, biotechnology, and medical applications, the patents on CRISPR-Cas9 are likely to be worth an enormous sum of money and worth fighting for.

In March of 2013, Jennifer Doudna (University of California, Berkeley) and colleagues filed the first patent applications for the use of an engineered CRISPR-Cas9 system. They claimed an invention date of May 25, 2012, a few months before their hallmark publication demonstrating that a single guide RNA could direct Cas9 to cut specific sequences *in vitro* [Jinek, M., et al. (2012). *Science* 337: 816–821].

However, in October of 2013, Feng Zhang of the Broad Institute filed patent applications for the use of CRISPR-Cas9 for genome editing in eukaryotic cells. Zhang claimed an invention date of December 12, 2012, based on his lab's seminal publication showing editing of cultured mammalian cells [Cong, L., et al. (2013). *Science* 339: 819–823].

Although Doudna and colleagues filed their applications first, Zhang also filed an "accelerated examination request," and paid a modest fee, to have his patents reviewed more quickly. Zhang's first CRISPR-Cas9 patent was approved on April 15, 2014. Doudna and colleagues filed an appeal claiming that their invention came first and that Zhang's patents overlap with their earlier filed patent. This initiated a "patent interference"—a proceeding by the USPTO to determine who shall be awarded the patents.

On December 6, 2016, lawyers representing each side met at the USPTO for a hearing. It is clear that Doudna and colleagues were the first to demonstrate cutting of specific sequences *in vitro* with a simplified CRISPR-Cas9 system. It is also clear that Zhang's lab was the first to show CRISPR-Cas9 editing of eukaryotic genomes *in vivo*. However, the central issue of the patent interference hearings was whether moving from an *in vitro* system to genome editing *in vivo* was an "obvious" next step with "reasonable expectations of success." To this point, lawyers for Doudna argued that six different labs reported success with CRISPR-Cas9 genome editing in eukaryotic cells within 6 months of their paper describing success *in vitro*. Lawyers for Zhang pointed to a 2012 interview with Doudna in which she expressed uncertainty about whether CRISPR-Cas would work in eukaryotic cells, thus seemingly undercutting claims to obviousness and setting a low expectation of success.

On February 15, 2017, the USPTO decided in favor of Zhang and the Broad Institute. Zhang maintains patents on CRISPR-Cas9 use in eukaryotic cells, while Doudna and colleagues maintain patents on CRISPR-Cas9 use broadly. What does this mean? Some legal experts believe the ruling is clearly in favor of Zhang and the Broad Institute. Others believe that Doudna would have a right to prosecute use of CRISPR-Cas9 in eukaryotic cells without her and UC Berkeley's approval, and that those using the technology in eukaryotic cells would require approval from both parties.

While the legal ramifications of the CRISPR-Cas9 patent battle are not yet clear, science is moving on. CRISPR-Cas9 clinical trials are under way, and Cas9 alternatives, such as Cpf1 from a different bacterial species, has some promising advantages over Cas9—such as a smaller size making it easier to deliver into cells. Zhang and colleagues hold the patent to CRISPR-Cpf1 technology in Europe and have patent applications pending in the United States. But surely other Cas9 alternatives are out there.

not yet been investigated. Even though the human genome has been completely sequenced, we are far from ascribing detailed functions for every gene. A simple way to learn about the function of a gene is to delete it and observe the consequences—an example of reverse genetics. Although this is conceptually simple, the ability to efficiently and quickly delete a gene from the genome has only recently become possible with CRISPR-Cas technology.

CRISPR-Cas9 is also an extremely versatile system. Mutations to the HNH and RuvC nuclease domains create a "dead" version of Cas9, known as **dCas9**, which can still bind but cannot cut target DNA. Using dCas9 as a platform, additional modifications have produced a versatile toolkit for research. For example, genetically fusing the activation domain of a transcriptional activator to dCas9 creates a custom tool that can be used to activate the transcription of any gene by simply designing an appropriate targeting sgRNA [**Figure ST 1.6(a)**]. Similarly, attaching the repression domain of a transcriptional repressor enables a researcher to turn off any given target gene [**Figure ST 1.6(b)**].

Genes can also be regulated at the epigenetic level (Chapter 19). By appending the catalytic domain of an epigenetic modifier enzyme, such as a histone deacetylase or a DNA methyltransferase, to dCas9, specific sequences

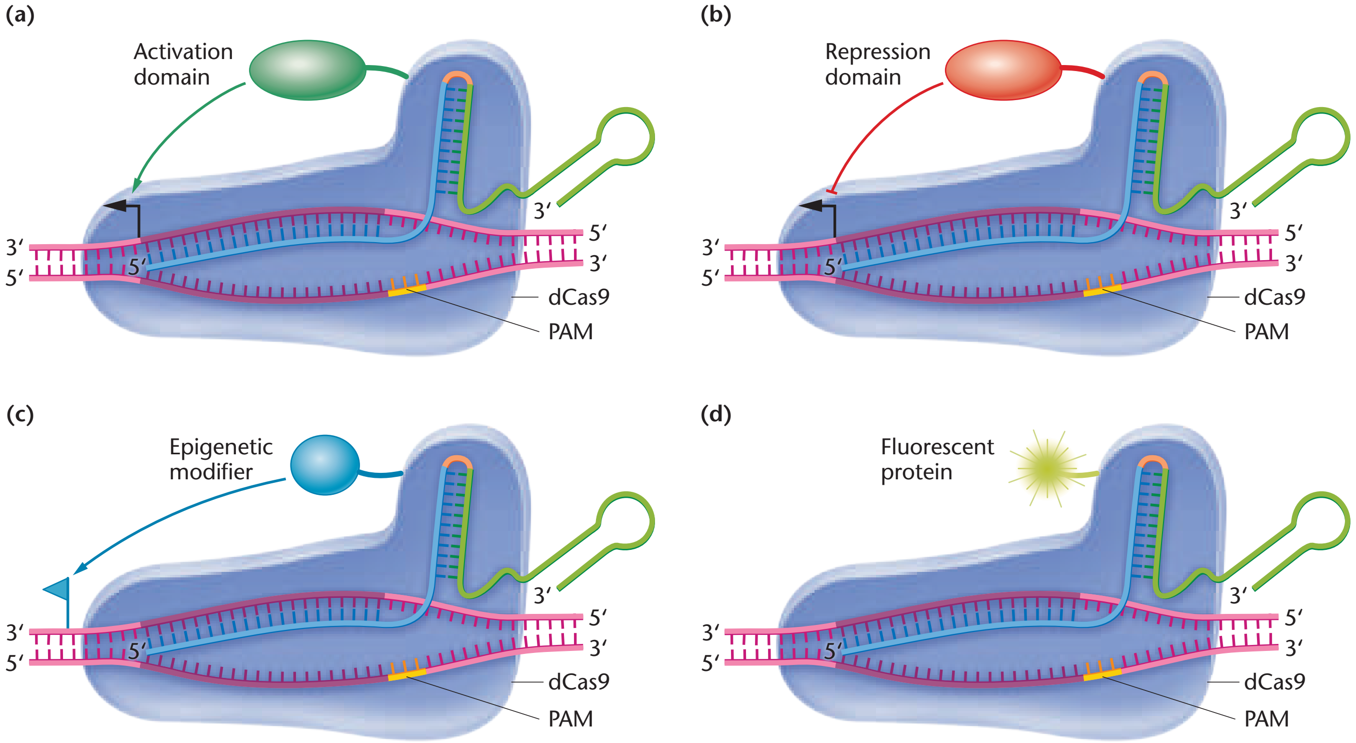


FIGURE ST 1.6 Modifications impart versatility to Cas9 as a tool for genetics research.

can be targeted for the addition or removal of epigenetic tags [**Figure ST 1.6(c)**]. In another approach, if fluorescent proteins, such as the green fluorescent protein (GFP), are attached to dCas9, a researcher can visually identify, using fluorescence microscopy, where a specific sequence is located within a cell [**Figure ST 1.6(d)**].

Recently, David Liu's lab has created a "base editor" by attaching a base-editing enzyme to dCas9. Cytidine deaminases (Chapter 13) convert the base cytosine into uracil, which is then replaced with thymine by the cell's endogenous base excision repair mechanism. By targeting a cytidine deaminase-dCas9 hybrid protein to a specific sequence with an sgRNA, a C-to-G change can be made. Since base editor technology does not induce cuts in the DNA, it may avoid the indels that are a common unwanted consequence of imprecise double-strand break repair.

Using the dCas9 toolkit, a researcher can turn the expression of genes on or off at the transcriptional or epigenetic levels, or create precise base-pair changes. When these manipulations are paired with observations of the effects on cells, tissues, or entire organisms, researchers can thoroughly investigate gene function. The applications of CRISPR-Cas technology for research are so vast that the primary limitation may well in fact be the imagination of the researcher.

CRISPR-Cas in Biotechnology

Biotechnology is the use of living organisms to create a product or a process that helps improve the quality of life for humans or other organisms (see Chapter 22). CRISPR-Cas has greatly facilitated biotechnological innovation because it enables the rapid and cost-effective production of genetically modified organisms for various purposes. CRISPR-Cas biotechnological applications range from simple but useful innovations, such as the creation of tomatoes that ripen more quickly, to massive and controversial endeavors, such as "bringing back" the woolly mammoth by editing mammoth genes into the elephant genome. The woolly mammoth is not back yet, but there are already many examples of CRISPR-Cas edited or modified organisms that serve biotechnological purposes.

A major challenge in raising livestock is managing disease. For example, each year the pig farming industry in the United States loses over $600 million to a single disease—porcine respiratory and reproductive syndrome (PRRS). PRRS results from infection by the porcine respiratory and reproductive syndrome virus (PRRSV). This virus infects cells of the immune system—macrophages—that reside in the pig's lungs, leading to respiratory complications and reproductive failure in pregnant sows. Studies have determined the PRRSV gains entry into macrophages via the cells' CD163 receptor. Therefore, a research team led by Randall Prather used CRISPR-Cas9 to remove the *CD163* gene from the pig genome. Pigs homozygous for *CD163* deletion showed no clinical signs of the disease following exposure to the virus. However, the CD163 receptor has important immune system functions such as mediating systemic inflammation and its deletion may cause immunological complications. To address this concern, a research team led by Alan Archibald engineered a precise modification of the *CD163* gene. The researchers used CRISPR-Cas9 to remove a small section of the gene that encodes a domain of the CD163 receptor that interacts with PRRSV. This more precise *CD163* edit yields pigs that are resistant to PRRSV but also retain the native immune system functions of the CD163 receptor.

CRISPR-Cas technology is currently also being used to modify food crops to introduce traits such as enhanced nutritional value, increased shelf life, and pest or drought resistance. For example, DuPont Pioneer®, a biotech firm specializing in hybrid crop seeds, used CRISPR-Cas to modify the *ARGOS8* gene in corn to create a drought-resistant strain. Past studies had shown that *ARGOS8* expression improves drought resistance. However, *ARGOS8* expression is low in corn. Therefore, scientists removed the native promoter of *ARGOS8* and introduced a promoter that directs stronger expression. The researchers inserted the promoter for the corn *GOS2* gene, a gene that encodes a translation factor that is constitutively expressed. Under drought conditions, *ARGOS8*-modified corn produced five bushels more per acre than unmodified corn. CRISPR-Cas has also been used to create mushrooms that resist browning after being sliced (see Special Topic Chapter 4—Genetically Modified Foods, for additional information on the use of CRISPR-Cas in creating gene-edited food crops). We are likely to see more CRISPR-Cas-derived foods in the grocery store soon.

CRISPR-Cas is also being used to produce human proteins in nonhuman animals for medical and research purposes. Albumin is the most abundant protein in blood plasma where it binds ions, hormones, and fatty acids. Patients with severe burns, traumatic shock, or liver disease are often treated with albumin. Albumin is also added to drugs and vaccines and is an ingredient in cell culture media. To address the need for albumin, researchers at the Beijing Proteome Research Center used CRISPR-Cas technology to create pigs expressing human albumin protein instead of the pig version of the protein. The researchers targeted Cas9 to the pig's albumin coding sequence and replaced it with a human albumin coding sequence. Many in the biotechnology field anticipate a surge in the use of CRISPR-Cas—edited large animals as bioreactors to produce human proteins for medical purposes.

Another biotechnology application of CRISPR-Cas has been the use of "gene drive"—a strategy to increase the chances that a specific allele becomes more prevalent in a population. (Refer to Chapter 22 for a Case Study on the ethical use of gene drive to control mosquito populations and combat the spread of mosquito-borne diseases.)

Clinical Use of CRISPR-Cas to Treat or Cure Disease

Perhaps one of the most anticipated applications of CRISPR-Cas technology is for treating or even curing, human genetic diseases—in other words, gene therapy. (For a broader discussion of gene therapy, see Special Topic Chapter 5—Gene Therapy.) The concept is simple: correct a disease-causing mutation in the genome to treat or cure the disease. However, there are many hurdles and considerations. For example, the specific mutation causing the disease must be known, and the tissue or cell type in which the disease manifests itself must be known and accessible. Furthermore, diseases that cause developmental defects during embryogenesis cannot be treated by correcting mutations after birth. There are also safety concerns: Can a disease-causing mutation be corrected without other unwanted changes to the genome? Finally, is it ethical to do so? See Box 2 for a discussion of the ethical use of CRISPR-Cas technology.

There are different strategies for treating recessive versus dominant diseases with CRISPR-Cas technology. For recessive disorders, CRISPR-Cas can be used to induce a double-stranded break in the gene bearing the disease-causing mutation, which can then be corrected with a donor template encoding the wild-type gene sequence. Although editing both alleles is possible, correction of a single allele is often sufficient. For dominant disorders, most afflicted individuals are heterozygous; they have a dominant mutant allele and a normal allele. CRISPR-Cas may be used to specifically inactivate the dominant mutant allele (with an allele-specific sgRNA), rendering it a recessive, loss-of-function allele. In many cases, the undisrupted allele would then be able to restore normal gene function.

So far, work with animal disease models and cultured human cells has provided a proof of principle that CRISPR-Cas can be used to correct disease-causing mutations in the genome. One promising study comes from a mouse model of Duchenne muscular dystrophy (DMD), a genetic disorder that leads to muscle degeneration, and eventually death. DMD is caused by mutations in the *DMD* gene, which encodes dystrophin, a cytoskeletal protein essential for muscle function. Previous work established a mouse model of DMD by introducing a stop codon into exon 23 of the mouse *Dmd* gene, thus leading to a truncated and nonfunctional protein. These mutant mice exhibit DMD-like muscle degeneration. To test the feasibility of using CRISPR-Cas to correct the *Dmd* mutation, Eric Olson's lab injected Cas9, an sgRNA, and a donor template carrying wild-type *Dmd* sequence into mutant

BOX 2

Ethical Concerns of Human Genome Editing

Should CRISPR-Cas technology be used to edit the human genome?

In December of 2015, roughly 500 scientists, ethicists, policy makers, and advocacy groups gathered in Washington, D.C. for a three-day summit to discuss this question and others.

Central to the discussion was the difference between editing the human genome in cultured cells for biomedical research, editing in human somatic (nonreproductive) cells for treating disease, and editing in germ cells (reproductive cells) or embryos to prevent disease in the next generation. Because editing of germ cells, or embryos, leads to changes that can be inherited by future generations, this type of genome editing carries the broadest implications. Some at the summit called for a moratorium on germ-line genome editing; others outlined plans to cure diseases with such technology.

Earlier in 2015, prior to the summit, researchers at Sun Yat-sen University in China were the first to use CRISPR-Cas to edit human embryos. Their goal was to edit the *HBB* gene to eliminate the β-thalassemia-causing mutation. However, the experiments were conducted with triploid embryos that could not survive, even if implanted. So controversial was this study that the high-profile journals *Science* and *Nature* refused to consider it for publication. Nonetheless, the work was published in *Protein & Cell*. The main findings of the study were that specific editing was possible, but with a low efficiency and with unintended edits. The authors concluded that CRISPR-Cas editing of human embryos is not yet safe.

With ongoing debate among scientists, ethicists, and broadly in the public, the National Academy of Sciences and the National Academy of Medicine convened a panel of 22 experts from several countries to ponder the science, ethics, and governance of human genome editing. They issued a report in February of 2017 including guidelines that purport to take into account the potential benefits versus unintended harms, societal values, and different perspectives across nations and cultures.

The panel concluded that somatic cell editing for therapeutic purposes should proceed. In addition, the panel recommended that the acceptable levels of editing efficiency be evaluated in the context of each intended application. However, the panel recognized that somatic cell editing in individuals with no pathology could be used for enhancement purposes, such as stronger muscles. The panel concluded that somatic cell editing for enhancement purposes should not proceed—at least not at this time.

The panel also considered germ-line editing. One benefit could be enabling prospective parents carrying disease alleles to have offspring without the disease, even if both parents are homozygous for a disease-causing mutation. However, this potential benefit weighs against the risk of unintended genome edits that may confer other problems. In addition, the panel put forth social and religious concerns that heritable genome edits are deemed unethical. Taking this all into account, the panel recommended caution, but not a ban on germ-line genome editing. Specifically, the panel recommended that germ-line editing "only be permitted for compelling reasons and under strict oversight."

Clearly, discussions of the potential benefits, safety concerns, and ethical issues related to editing the human genome will continue. In August of 2017, a team led by Shoukhrat Mitalipov generated viable human embryos (using an in vitro technique known as intracytoplasmic sperm injection) and then injected them with CRISPR-Cas machinery to correct a dominant a mutation in the *MYBPC3* gene associated with heart disease and sudden death. The embryos were not implanted and were only allowed to progress to the blastocyst stage (~5 days). Although the study reported successful editing, others are skeptical and additional studies are needed to determine the efficacy of editing human embryos. Results aside, is it ethical to create human embryos specifically for genome-editing experiments? In the same month, a group led by Kathy Niakan edited human embryos (again not implanted) to determine the role of the *POU5F1* gene in embryonic development. This study used unwanted zygotes donated from an in vitro fertilization (IVF) clinic. Is it ethical to perform genome-editing experiments on IVF "leftovers"?

Currently, the United States prohibits the use of federal funds to modify a human embryo. However, the Mitalipov study was privately funded and the Niakan lab is in the United Kingdom. Across the globe, the legality of editing human embryos varies widely. Some countries have banned the editing of human embryos, some have set some restrictions, and others yet have no restrictions at all.

mouse zygotes. The researchers found that injected zygotes produced mosaic mice—some mouse cells were edited, while others were not. Nonetheless, even mice with a small number of edited cells exhibited significant rescue of muscle defects.

Although this success in a mouse model is promising, the strategy of injecting embryos is impossible for human DMD patients that are typically diagnosed at the age of 4 when symptoms begin. Delivery of CRISPR-Cas machinery to muscle cells in a 4 year old is a far more difficult challenge than delivery to a single-celled zygote. Yet, it may be possible. A 2017 study led by Irina Conboy and Niren Murthy showed that a strategy called "CRISPR-Gold" has remarkable promise as as CRISPR-Cas delivery vehicle.

Gold nanoparticles were labeled with Cas9, sgRNAs, donor templates, and an endosomal disruptive polymer. When injected into tissue, CRISPR-Gold is endocytosed (internalized) by the cells and becomes contained within membrane-bound endosomes. However, the endosomal disruptive polymer breaks open the endosomes, thus releasing the editing machinery. Intramuscular injection of CRISPR-Gold designed for *Dmd* repair into 8-week-old *Dmd* mutant mice led to significant dystrophin protein expression and greatly improved muscle function within two weeks!

In 2016, a team led by David Martin, Dana Carroll, and Jacob Corn moved a step closer to clinical application with CRISPR-Cas. The researchers extracted bone marrow stem cells from patients with sickle-cell anemia and then used CRISPR-Cas to correct the disease-causing mutation in the β-globin-encoding gene (*HBB*) in the extracted cells. Bone marrow stem cells are precursors of red blood cells, which are affected in patients with sickle-cell anemia. When the edited stem cells differentiated into red blood cells, they expressed wild-type β-globin. Next, the researchers grafted the edited stem cells into mice. (The mice were immunocompromised to prevent rejection.) Even 16 weeks after the graft, edited red blood cells expressing wild-type β-globin were detected in the mice, suggesting that the effects could be long lasting.

While editing the genome of bone marrow stem cells may one day be used to alleviate the symptoms of a sickle-cell anemia patient, it may eventually be possible to edit human embryos and correct disease-causing mutations before any symptoms manifest. Some studies have already attempted CRISPR-Cas editing of human embryos to evaluate the efficacy, specificity, and safety of correcting disease-causing mutations. In 2015, a research group from Sun Yat-sen University in China showed that editing of the *HBB* gene in human embryos was possible, but that the efficiency was low, there were unwanted off-target edits, and the embryos were mosaic with only some cells carrying edits. In 2016, another group from China reported similar results for a different target gene. Both studies used triploid, and thus inviable embryos to assuage ethical concerns (see Box 2).

While editing of human embryos is not yet ready for medical application, clinical trials of CRISPR-Cas edited somatic cells have already begun. In October of 2016, a research team led by Lu You of Sichuan University in China initiated a CRISPR-Cas–based clinical trial of non–small-cell lung cancer treatment for patients who had not had success with other treatments. The procedure involves extracting T cells, a type of immune cell, from the patient, inactivating the *PD-1* gene in these cells, and reintroducing the modified cells into the patient. PD-1 protein is a negative regulator of the immune response, and removing it may enable the immune system to attack the cancer more aggressively. The trial aims to enroll 10 patients to determine if there are any adverse side effects of the treatment and to determine its efficacy. As of August 2017, three other clinical trials, all in China, were recruiting patients to use *PD-1* knockout T cells to treat other types of cancer. A similar clinical trial for treating cancer is set to begin in the United States in 2017 but with edits to three different genes in T cells. (For more information on similar immunotherapies based on gene-modified T cells, see Special Topic Chapter 3—Genomics and Precision Medicine).

Numerous other CRISPR-Cas–based clinical trials are being planned for a wide variety of genetic disorders such as muscle, blood, and liver diseases, as well as heritable blindness, HIV infection, and various cancers.

Selected Readings and Resources

Journal Articles

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau S., Romero, D. A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas Systems. *Science* 339:819–823.

Fogarty et al., (2017). https://www.nature.com/articles/nature24033

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. USA* 109(39):E2579–E2586.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* 337:816–821.

Lee et al., (2017). https://www.nature.com/articles/s41551-017-0137-2

Marraffini, L. A., and Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in Staphylococci by targeting DNA. *Science* 322:1843–1845.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science* 339:823–826.

Pennisi, E. (2013). The CRISPR craze. *Science* 341:833–836.

Urnov, F. (2016) Genome editing: The domestication of Cas9. *Nature* 529:468–469.

Xiong, X., Chen, M., Lim, W. A., Zhao, D., and Qi, L. S. (2016). CRISPR/Cas9 for human genome engineering and disease research. *Annu. Rev. Genom. Hum. Genet.* 17:131–154.

Web Sites

Broad Institute: Research Highlights: CRISPR. https://www.broadinstitute.org/research-highlights-crispr

Caribou Biosciences. http://cariboubio.com/

CRISPR : Nature News & Comment. http://www.nature.com/news/crispr-1.17547

CRISPR Therapeutics. http://crisprtx.com/

Editas Medicine. http://www.editasmedicine.com/

iBiology: Genome Engineering with CRISPR-Cas9: Birth of a Breakthrough Technology. https://www.ibiology.org/ibiomagazine/jennifer-doudna-genome-engineering-with-crispr-cas9-birth-of-a-breakthrough-technology.html

National Academy of Sciences and National Academy of Medicine: Report: Human Genome Editing—Science, Ethics, and Governance http://nationalacademies.org/cs/groups/genesite/documents/ webpage/gene_177260.pdf

Review Questions

1. What is the difference between innate immunity and adaptive immunity in bacteria?
2. What evidence demonstrates that CRISPR-Cas is an adaptive immune system defense against viruses?
3. Describe the three steps of CRISPR-Cas–mediated antiviral defense.
4. Why was the type II CRISPR-Cas9 system of *S. pyogenes* selected by several scientists as a model to learn how the CRISPR-Cas mechanism works?
5. How does the CRISPR-Cas system of *S. pyogenes* distinguish "self" DNA from foreign DNA?
6. What is a single guide RNA, and what role does it play in CRISPR-Cas genome editing in eukaryotic cells?
7. What is the difference between nonhomologous end-joining (NHEJ) and homology-directed repair (HDR) in the context of genome editing?
8. Provide one example of a CRISPR-Cas application for biotechnology.
9. How must recessive and dominant mutations be treated differently, in theory, when correcting disease mutations with CRISPR-Cas?

Discussion Questions

1. What advantages does CRISPR-Cas have over previously used genome-editing technologies such as zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs)?
2. Based on pure chance alone, how often would you expect to find a perfect match in the human genome to a 20-nucleotide sgRNA targeting sequence that is adjacent to a PAM sequence? The human haploid genome has 3.2 billion base pairs.
3. What ethical and safety considerations must be taken before CRISPR-Cas is used to edit human embryos to cure disease or to engineer desirable traits?
4. Recall (from Chapter 18) how miRNAs and the RNA-induced silencing complex (RISC) regulate gene expression in eukaryotes. What about that system is conceptually similar to the CRISPR-Cas system? What is conceptually different?
5. Describe two different ways in which engineered mutations of Cas9 have led to improved utility for genome editing.
6. Consider the following human genetic diseases: hemophilia, Down syndrome, cystic fibrosis, and brain cancer. Which are the best candidates for treatment with CRISPR-Cas genome editing, and which have the largest hurdles to overcome? Why?
7. What are the different concerns about off-target edits by CRISPR-Cas when editing the human genome in embryos versus adult somatic cells?

DNA Forensics

Forensic science (or *forensics)* uses technological and scientific approaches to answer questions about the facts of criminal or civil cases. Prior to 1986, forensic scientists had a limited array of tools with which to link evidence to specific individuals or suspects. These included some reliable methods such as blood typing and fingerprint analysis, but also many unreliable methods such as bite mark comparisons and hair microscopy.

Since the first forensic use of **DNA profiling** in 1986 (Box 1), **DNA forensics** (also called **forensic DNA fingerprinting** or **DNA typing)** has become an important method for police to identify sources of biological materials. DNA profiles can now be obtained from saliva left on cigarette butts or postage stamps, pet hairs found at crime scenes, or bloodspots the size of pinheads. Even biological samples that are degraded by fire or time are yielding DNA profiles that help the legal system determine identity, innocence, or guilt. Investigators now scan large databases of stored DNA profiles in order to match profiles generated from crime scene evidence. DNA profiling has proven the innocence of people who were convicted of serious crimes and even sentenced to death. Forensic scientists have used DNA profiling to identify victims of mass disasters such as the Asian Tsunami of 2004 and the September 11, 2001, terrorist attacks in New York. They have also used forensic DNA analysis to identify endangered species and animals trafficked in the illegal wildlife trade.

"Even biological samples degraded by fire or time are yielding DNA profiles that help determine identity, innocence, or guilt."

The applications of DNA profiling extend beyond forensic investigations. These include paternity and family relationship testing, identification of plant materials, verification of military casualties, and evolutionary studies.

It is important to understand the basics of forensic DNA analysis. As informed citizens, we need to monitor its uses and potential misuses. Although DNA profiling is well validated as a technique and is considered the gold standard of forensic identification, it is not without controversy and the need for legislative oversight.

In this Special Topic chapter, we will explore how DNA profiling works and how the results of profiles are interpreted. We will learn about DNA databases, the potential problems associated with DNA profiling, and the future of this powerful technology.

ST 2.1 DNA Profiling Methods

VNTR-Based DNA Fingerprinting

The era of DNA-based human identification began in 1985, with Dr. Alec Jeffreys's publication on DNA loci known as **minisatellites,** or **variable number of tandem repeats (VNTRs).** As described earlier in the text (see Chapter 12), VNTRs are located in noncoding regions of the genome and are made up of DNA sequences of between 15 and 100 base pairs (bp) long, with each unit repeated a number of times. The number of repeats found at each VNTR locus varies from person to person, and hence VNTRs can be from 1 to 20 kilobases (kb) in length, depending on the person. For example, the VNTR

5′- GACTGCCTGCTAAGAT**GACTGCCTGCTAAGAT**GACTGCCTGCTAAGAT-3′

is composed of three tandem repeats of a 16-nucleotide sequence (highlighted in bold).

VNTRs are useful for DNA profiling because there are as many as 30 different possible alleles (repeat lengths) at any VNTR in a population. This creates a large number of possible genotypes. For example, if one examined four different VNTR loci within a population, and each locus had 20 possible alleles, there would be more than 2 billion (4^{20}) possible genotypes in this four-locus profile.

To create a VNTR profile (also known as a DNA fingerprint), scientists extract DNA from a tissue sample and digest it with a restriction enzyme that cleaves on either side of the VNTR repeat region (**Figure ST 2.1**). The digested DNA is separated by gel electrophoresis and subjected to Southern blot analysis (which is described in detail in Chapter 20). Briefly, separated DNA is transferred from the gel to a

BOX 1

The Pitchfork Case: The First Criminal Conviction Using DNA Profiling

In the mid-1980s, the bodies of two schoolgirls, Lynda Mann and Dawn Ashworth, were found in Leicestershire, England. Both girls had been raped, strangled, and their bodies left in the bushes. In the absence of useful clues, the police questioned a local mentally retarded porter named Richard Buckland. During interrogation, Buckland confessed to the murder of Dawn Ashworth; however, police did not know whether he was also responsible for Lynda Mann's death. In 1986, in order to identify the second killer, the police asked Dr. Alec Jeffreys of the University of Leicester to analyze the crime scene evidence using a new method of DNA analysis called VNTR profiling. Dr. Jeffreys's VNTR analysis revealed a match between the DNA profiles from semen samples obtained from both crime scenes, suggesting that the same person was responsible for both rapes. However, neither of the DNA profiles matched those from a blood sample taken from Richard Buckland. Having eliminated their only suspect, the police embarked on the first mass DNA dragnet in history by requesting blood samples from every adult male in the region. Although 4000 men offered samples, one did not. Colin Pitchfork, a bakery worker, paid a friend to give a blood sample in his place, using forged identity documents. Their plan was detected when their conversation was overheard at a local pub. The conversation was reported to police, who then arrested Pitchfork, obtained his blood sample, and sent it for analysis. His DNA profile matched the profiles from the semen samples left at both crime scenes. Pitchfork confessed to the murders, pleaded guilty, and was sentenced to life in prison. The Pitchfork Case was not only the first criminal case resolved by forensic DNA profiling, but also the first case in which DNA profiling led to the exoneration of an innocent person.

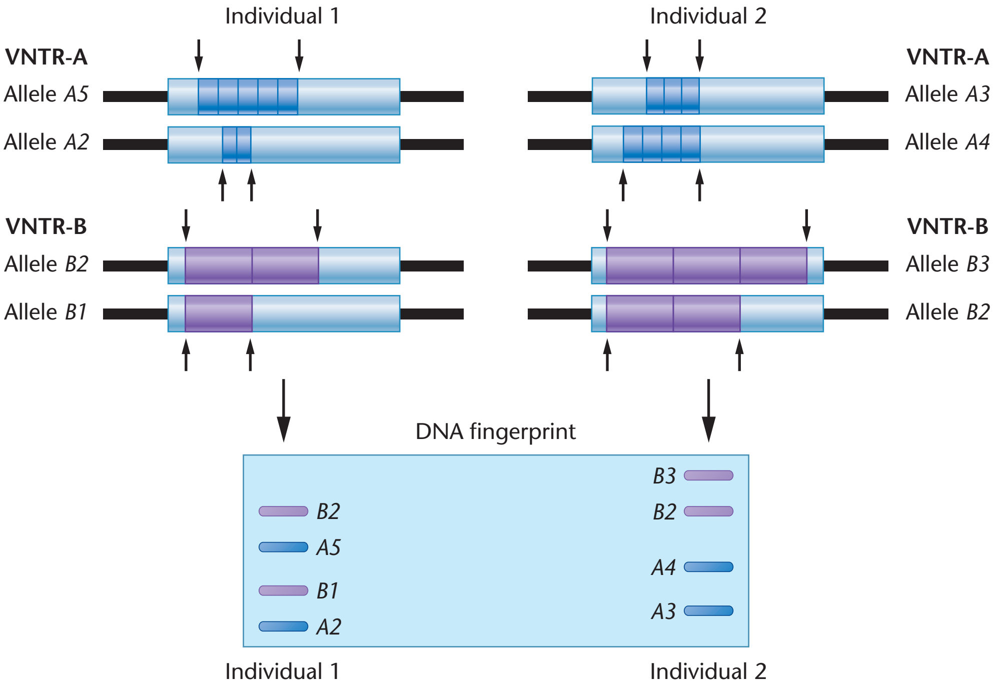


FIGURE ST 2.1 DNA fingerprint at two VNTR loci for two individuals. VNTR alleles at two loci (*A* and *B*) are shown for two different individuals. Arrows mark restriction-enzyme cutting sites that flank the VNTRs. Restriction-enzyme digestion produces a series of fragments that can be separated by gel electrophoresis and detected as bands on a Southern blot (bottom). The number of repeats at each locus is variable, so the overall pattern of bands is distinct for each individual. The DNA fingerprint profile shows that these individuals share one allele (*B2*).

membrane and hybridized with a radioactive probe that recognizes DNA sequences within the VNTR region. After exposing the membrane to X-ray film, the pattern of bands is measured, with larger VNTR repeat alleles remaining near the top of the gel and smaller VNTRs, which migrate more rapidly through the gel, being closer to the bottom. The pattern of bands is the same for a given individual, no matter what tissue is used as the source of the DNA. If enough VNTRs are analyzed, each person's DNA profile will be unique (except, of course, for identical twins) because of the huge number of possible VNTRs and alleles. In practice, scientists analyze about five or six loci to create a DNA profile.

A significant limitation of VNTR profiling is that it requires a relatively large sample of DNA (10,000 cells or about 50 μg of DNA)—more than is usually found at a typical crime scene. In addition, the DNA must be relatively intact (nondegraded). As a result, VNTR profiling has been used most frequently when large tissue samples are available—such as in paternity testing. Although VNTR profiling is still used in some cases, it has mostly been replaced by more sensitive methods, as described next.

Autosomal STR DNA Profiling

The development of the **polymerase chain reaction (PCR)** revolutionized DNA profiling. As described in Chapter 20, PCR is an *in vitro* method that uses specific primers and a heat-tolerant DNA polymerase to amplify specific regions of DNA. Within a few hours, this method can generate a millionfold increase in the quantity of DNA within a specific sequence region. Using PCR-amplified DNA samples, scientists are able to generate DNA profiles from trace samples (e.g., the bulb of single hairs or a few cells from a bloodstain) and from samples that are old or degraded (such as a bone found in a field or an ancient Egyptian mummy).

The majority of human forensic DNA profiling is now done by amplifying and analyzing regions of the genome known as **microsatellites,** or **short tandem repeats (STRs).** STRs are similar to VNTRs, but the repeated motif is shorter—between two and nine base pairs, repeated from 7 to 40 times. For example, one locus known as D8S1179 is made up of the four base-pair sequence TCTA, repeated 7 to 20 times, depending on the allele. There are 19 possible alleles of the locus that are found within a population. Although hundreds of STR loci are present in the human genome, only a subset is used for DNA profiling. At the present time, the FBI and other U.S. law enforcement agencies use 20 STR loci as a core set for forensic analysis. Most European countries now use 12 STR loci as a core set.

Several commercially available kits are used for forensic DNA analysis of STR loci. The methods vary slightly, but generally involve the following steps. As shown in **Figure ST 2.2**, each primer set is tagged by one of four fluorescent dyes—represented here as blue, green, yellow, or red. Each primer set is designed to amplify DNA fragments, the sizes of which vary depending on the number of repeats within the region amplified. For example, the primer sets that amplify the TH01, vWA, D21S11, D7S820, D5S818, TPOX, and DYS391 STR loci are all labeled with a fluorescent tag indicated as yellow. The sizes of the amplified DNA fragments allow scientists to differentiate between the yellow-labeled products. For example, the amplified products from the D21S11 locus range from about 200 to 260 bp in length, whereas those from the TPOX locus range from about 375 to 425 bp, and so on.

After amplification, the DNA sample will contain a small amount of the original template DNA sample and a large amount of fluorescently labeled amplification products (**Figure ST 2.3**). The sizes of the amplified fragments are measured by **capillary electrophoresis.** This method uses thin glass tubes that are filled with a polyacrylamide gel material similar to that used in slab gel electrophoresis. The amplified DNA sample is loaded onto the top of the capillary tube, and an electric current is passed through

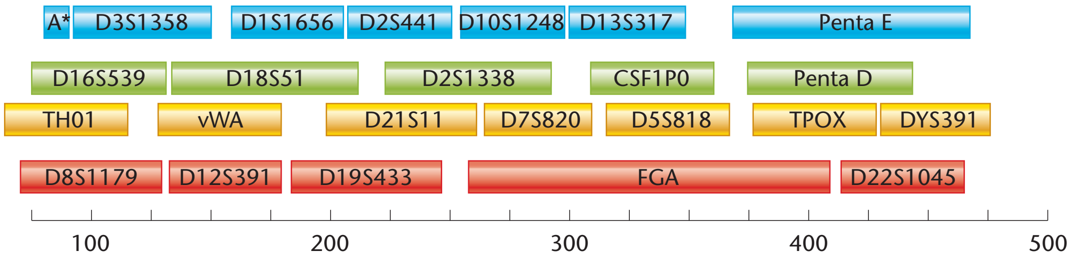


FIGURE ST 2.2 Relative size ranges and fluorescent dye labeling colors of 24 STR products generated by a commercially available DNA profiling kit. The scale at the bottom of the diagram indicates DNA fragment sizes in base pairs.

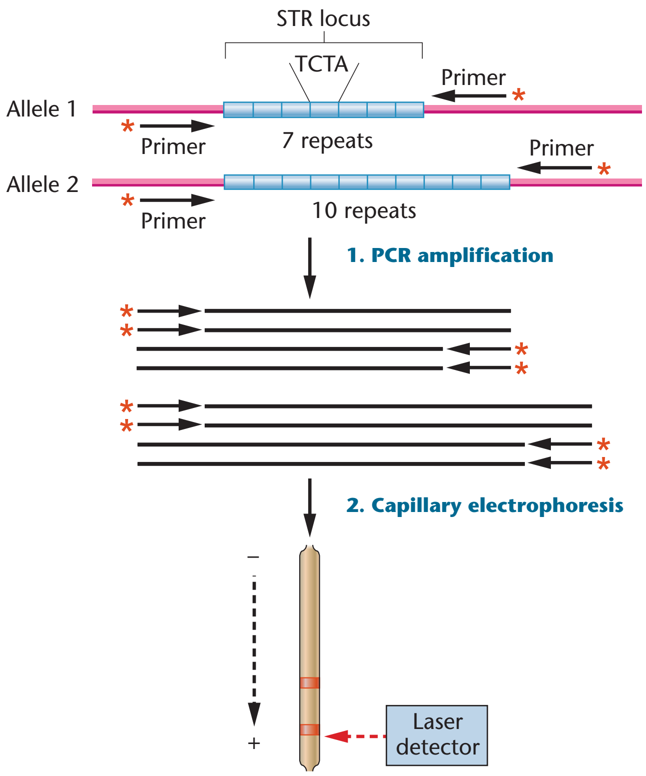


FIGURE ST 2.3 Steps in the PCR amplification and analysis of one STR locus (D8S1179). In this example, the person is heterozygous at the D8S1179 locus: One allele has 7 repeats and one has 10 repeats. Primers are specific for sequences flanking the STR locus and are labeled with a red fluorescent dye. The double-stranded DNA is denatured, the primers are annealed, and each allele is amplified by PCR in the presence of all four dNTPs and Taq DNA polymerase. After amplification, the labeled products are separated according to size by capillary electrophoresis, followed by fluorescence detection.

the tube. The negatively charged DNA fragments migrate through the gel toward the positive electrode, according to their sizes. Short fragments move more quickly through the gel, and larger ones more slowly. At the bottom of the tube, a laser detects each fluorescent fragment as it migrates through the tube. The data are analyzed by software that calculates both the sizes of the fragments and their quantities, and these are represented as peaks on a graph (Figure ST 2.4). Typically, automated systems analyze dozens of samples at a time, and the analysis takes less than an hour.

After DNA profiling, the profile can be directly compared to a profile from another person, from crime scene evidence, or from other profiles stored in DNA profile databases (Figure ST 2.5). The STR profile genotype of an individual is expressed as the number of times the STR sequence is repeated. For example, in the profile shown in Figure ST 2.5, the person's profile would be expressed as shown in **Table ST 2.1**.

Scientists interpret STR profiles using statistics, probability, and population genetics, and these methods will be discussed in the section Interpreting DNA Profiles.

Y-Chromosome STR Profiling

In many forensic applications, it is important to differentiate the DNA profiles of two or more people in a mixed sample. For example, vaginal swabs from rape cases usually contain a mixture of female somatic cells and male sperm cells. In addition, some crime samples may contain evidence material from a number of male suspects. In these types of cases, STR profiling of Y-chromosome DNA is useful. There are more than 200 STR loci on the Y chromosome that are useful for DNA profiling; however, fewer than 20 of these are used routinely for forensic analysis. PCR amplification of Y-chromosome STRs uses specific primers that do not amplify DNA on the X chromosome.

One limitation of Y-chromosome DNA profiling is that it cannot differentiate between the DNA from fathers and sons or from male siblings. This is because the Y chromosome is directly inherited from the father to his sons, as a single unit. The Y chromosome does not undergo recombination, meaning that less genetic variability exists on the Y chromosome than on autosomal chromosomes. Therefore, all patrilineal relatives share the same Y-chromosome profile. Even two apparently unrelated males may share the same Y profile, if they also share a distant male ancestor.

Although these features of Y-chromosome profiles present limitations for some forensic applications, they are useful for identifying missing persons when a male relative's DNA is available for comparison. They also allow researchers to trace paternal lineages in genetic genealogy studies.

Mitochondrial DNA Profiling

Another important addition to DNA profiling methods is **mitochondrial DNA (mtDNA)** analysis. Between 200 and 1700 mitochondria are present in each human somatic cell. Each mitochondrion contains one or more 16-kb circular DNA chromosomes. Mitochrondria are passed from the human egg cell to the zygote during fertilization; however, as sperm cells contribute few if any mitochondria to the zygote, they do not contribute these organelles to the next generation. Therefore, all cells in an individual contain multiple copies of specific mitochondrial variants derived from the mother. Like Y-chromosome DNA, mtDNA undergoes little if any recombination and is inherited as a single unit.

Scientists create mtDNA profiles by amplifying regions of mtDNA that show variability between unrelated individuals and populations. After PCR amplification, the DNA sequence within these regions is determined by automated

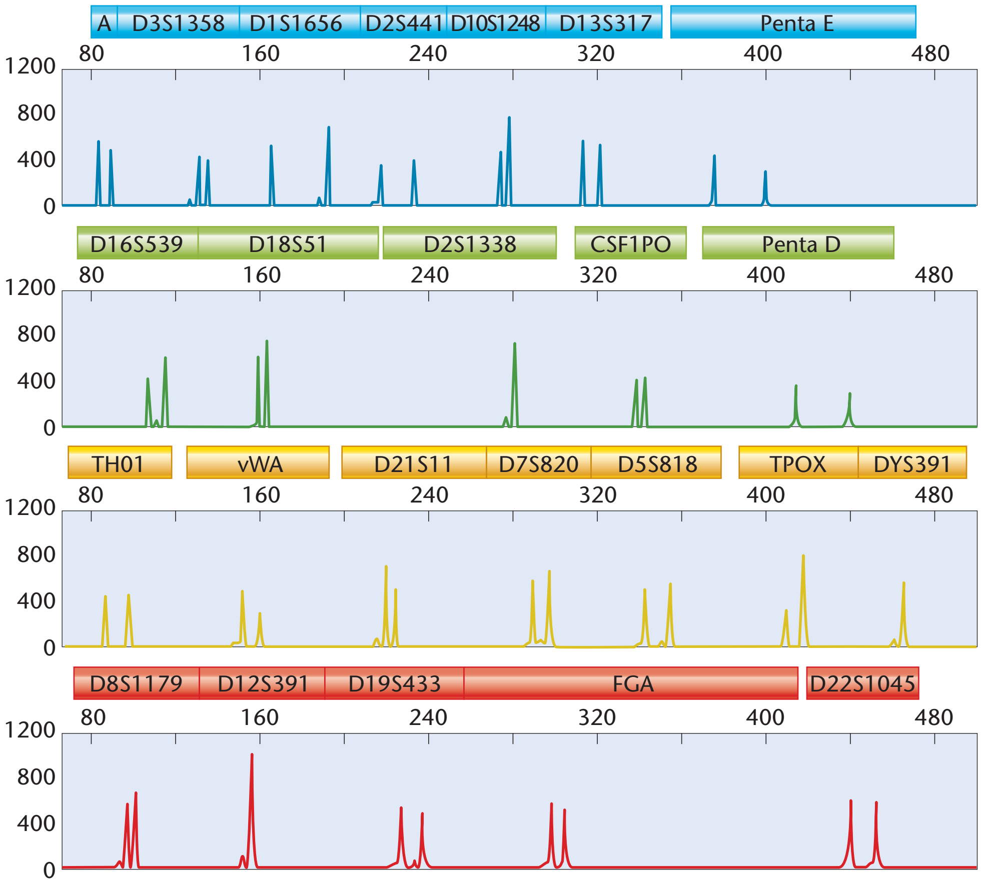


FIGURE ST 2.4 An electropherogram showing the results of a DNA profile analysis using the 24-locus STR profile kit shown in Figure ST 2.2. Heterozygous loci show up as double peaks, and homozygous loci as single, higher peaks. The sizes of each allele can be calculated from the peak, locations relative to the size axis shown at the top of each panel.

DNA sequencing. Scientists then compare the sequence with sequences from other individuals or crime samples, to determine whether or not they match.

The fact that mtDNA is present in high copy numbers in cells makes its analysis useful in cases where samples are small, old, or degraded. mtDNA profiling is particularly useful for identifying victims of mass murders or disasters, such as the Srebrenica massacre of 1995 and the World Trade Center attacks of 2001, where reference samples from relatives are available. The main disadvantage of mtDNA profiling is that it is not possible to differentiate

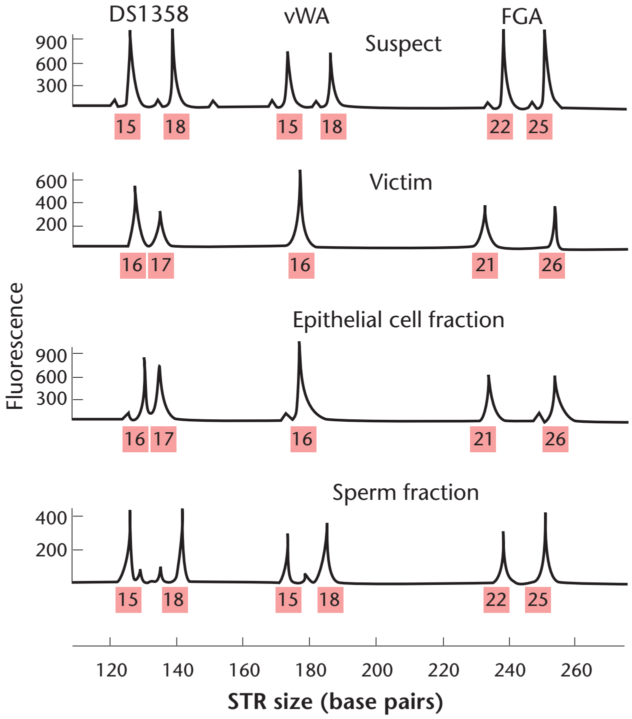


FIGURE ST 2.5 An electropherogram showing the STR profiles of four samples from a rape case. Three STR loci were examined from samples taken from a suspect, a victim, and two fractions from a vaginal swab taken from the victim. The x-axis shows the DNA size ladder, and the y-axis indicates relative fluorescence intensity. The number below each allele indicates the number of repeats in each allele, as measured against the DNA size ladder. Notice that the STR profile of the sperm sample taken from the victim matches that of the suspect.

between the mtDNA from maternal relatives or from siblings. Like Y-chromosome profiles, mtDNA profiles may be shared by two apparently unrelated individuals who also share a distant ancestor—in this case a maternal ancestor. Researchers use mtDNA profiles in scientific studies of genealogy, evolution, and human population migrations.

Single-Nucleotide Polymorphism Profiling

Single-nucleotide polymorphisms (SNPs) are single-nucleotide differences between two DNA molecules. They may be base-pair changes or small insertions or deletions (**Figure ST 2.6**). SNPs occur randomly throughout the genome, approximately every 500 to 1000 nucleotides. This means that there are potentially millions of loci in the human genome that can be used for profiling. However, as SNPs usually have only two alleles, many SNPs (50 or more) must be used to create a DNA profile that can distinguish between two individuals as efficiently as STRs.

Scientists analyze SNPs by using specific primers to amplify the regions of interest. The amplified DNA regions are then analyzed by a number of different methods such as automated DNA sequencing or hybridization to immobilized probes on DNA microarrays that distinguish between DNA molecules with single-nucleotide differences.

Forensic SNP profiling has one major advantage over STR profiling. Because a SNP involves only one nucleotide of a DNA molecule, the theoretical size of DNA required for a PCR reaction is the size of the two primers and one more nucleotide (i.e., about 50 nucleotides). This feature makes SNP analysis suitable for analyzing DNA samples that are severely degraded. Despite this advantage, SNP profiling has not yet become routine in forensic applications. More frequently, researchers use SNP profiling of Y-chromosome and mtDNA loci for lineage and evolution studies.

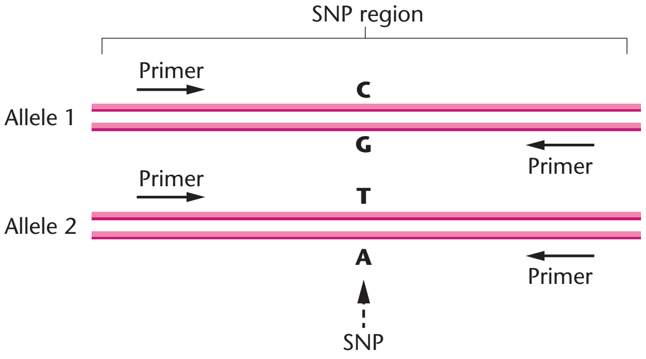


FIGURE ST 2.6 Example of a single-nucleotide polymorphism (SNP) from an individual who is heterozygous at the SNP locus. The arrows indicate the locations of PCR primers used to amplify the SNP region, prior to DNA sequence analysis. If this SNP locus only had two known alleles—the C and T alleles—there would be three possible genotypes in the population: CC, TT, and CT. The individual in this example has the CT genotype.

TABLE ST 2.1 **STR Profile Genotypes from the Four Profiles Shown in Figure ST 2.5**

	Profile Genotype from			
STR Locus	Suspect	Victim	Epithelial Cells	Sperm Fraction
DS1358	15, 18	16, 17	16, 17	15, 18
vWA	15, 18	16, 16	16, 16	15, 18
FGA	22, 25	21, 26	21, 26	22, 25

DNA Phenotyping

An emerging and controversial method, known as *DNA phenotyping*, is gaining popularity as a new DNA forensics tool. Unlike DNA profiling, which is used to confirm or exclude sample identities, DNA phenotyping uses DNA sequence information to reveal a person's physical features and ancestral origins.

Currently, DNA phenotyping methods can predict a person's eye, hair, and skin colors based on their DNA SNP patterns. For example, scientists have found six SNPs in six genes that are related to blue and brown eye color. Using statistical models based on these six SNPs, it is possible to predict with 95 percent accuracy whether a person has brown or blue eyes. Using 22 SNPs associated with 11 genes, it is possible to predict with 90 percent accuracy whether a person has black hair and 80 percent accuracy whether a person has red or brown hair. Skin color predictions involve 36 SNPs associated with 15 genes, with prediction accuracies similar to those for hair colors. Both biological sex and geographic ancestry can also be accurately determined from a person's DNA sequence.

Some researchers and private companies have taken DNA phenotyping well beyond prediction of these features. Their algorithms claim to predict 3-dimensional facial structures which allow them to compile full-color photographic representations of a person's face, based only on their DNA sample (**Figure ST 2.7**).

At the present time, DNA phenotyping has not been validated sufficiently to be presented in court. However, police are using the method to help identify unknown missing persons and to provide leads in cold cases (see Box 2). DNA phenotyping has raised concerns that it may lead to privacy violations and racial profiling. Scientists question its accuracy and reliability. Private companies specializing in DNA phenotyping do not reveal their methods, making it more difficult to validate their results. Many of these concerns may resolve in the future as this new technology becomes more sophisticated.

BOX 2

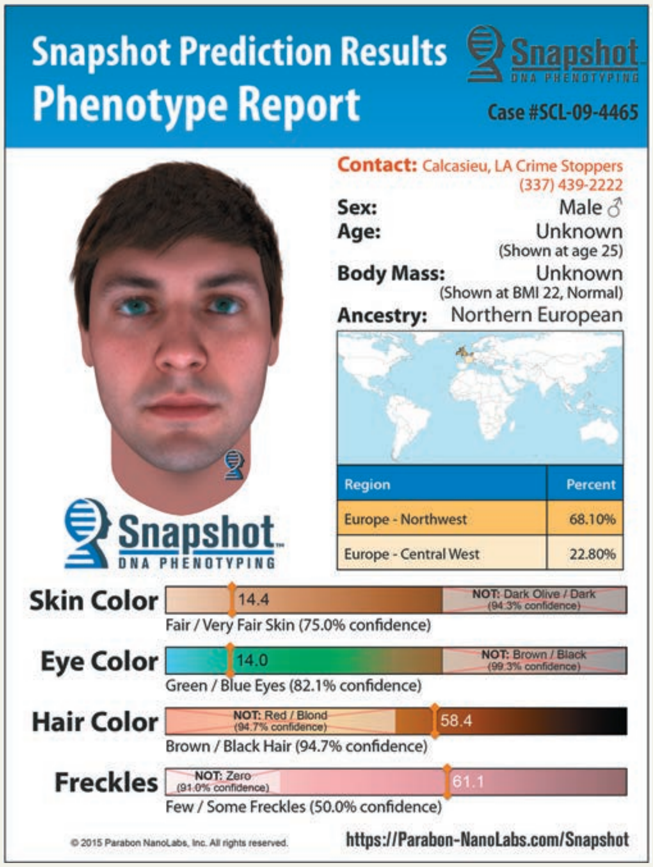
Putting a Face to DNA: The Bouzigard Case

On November 23, 2009, the body of 19-year-old Sierra Bouzigard was found near a rural road near Lake Charles, Louisiana. She had been beaten beyond recognition, but police were able to track her identity from her unique tattoo. They also isolated tissue deposited under her fingernails during her struggle with the murderer. DNA analysis gave a clear profile of a male, but no further information.

There were no witnesses to the crime, but police discovered that Bouzigard had been in the company of a crew of undocumented Mexican workers on the night she disappeared. Despite DNA analysis of swabs from each of the Mexican workers and a search of the CODIS DNA profile database, police were not able to find a match to the suspect's DNA. Bouzigard's family offered a $10,000 reward, and police requested information from the public. The police continued to look for a Hispanic male; however, in the absence of further leads, the case soon went cold.

In June of 2015, a major break in the Bouzigard case occurred. A DNA analyst working with the police department heard about DNA phenotyping and sent the suspect's DNA sample to Parabon NanoLabs, a company that specializes in generating DNA phenotype images. To the surprise of the investigators, the DNA phenotype analysis revealed that the suspect was a male with pale skin, blue or green eyes, brown hair, and freckles. He was of northern European ancestry (**Figure ST 2.7**).

With this new information, the police redirected their investigation and posted the DNA phenotype image to the media. A subsequent tip led to the arrest of a suspect. The suspect's DNA profile was found to match DNA profiles from samples present at the crime scene.



Snapshot Prediction Results
Phenotype Report

Snapshot DNA PHENOTYPING

Case #SCL-09-4465

Contact: Calcasieu, LA Crime Stoppers (337) 439-2222

Sex: Male ♂

Age: Unknown (Shown at age 25)

Body Mass: Unknown (Shown at BMI 22, Normal)

Ancestry: Northern European

Region	Percent
Europe - Northwest	68.10%
Europe - Central West	22.80%

Skin Color: 14.4 — NOT: Dark Olive / Dark (94.3% confidence) — Fair / Very Fair Skin (75.0% confidence)

Eye Color: 14.0 — NOT: Brown / Black (99.3% confidence) — Green / Blue Eyes (82.1% confidence)

Hair Color: NOT: Red / Blond (94.7% confidence) — 58.4 — Brown / Black Hair (94.7% confidence)

Freckles: NOT: Zero (91.0% confidence) — 61.1 — Few / Some Freckles (50.0% confidence)

© 2015 Parabon NanoLabs, Inc. All rights reserved.

https://Parabon-NanoLabs.com/Snapshot

FIGURE ST 2.7 A report summary of the Parabon® Snapshot™ analysis performed for the Calcasieu Parish Sheriff's Office, Louisiana, in support of their investigation into the 2009 murder of Sierra Bouzigard.

ST 2.2 Interpreting DNA Profiles

After a DNA profile is generated, its significance must be determined. In a typical forensic investigation, a profile derived from a suspect is compared to a profile from an evidence sample or to profiles already present in DNA databases. If the suspect's profile does not match that of the evidence profile or database entries, investigators can conclude that the suspect is not the source of the sample(s) that generated the other profile(s). However, if the suspect's profile matches the evidence profile or a database entry, the interpretation becomes more complicated. In this case, one could conclude that the two profiles either came from the same person—or they came from two different people who share the same DNA profile by chance. To determine the significance of any DNA profile match, it is necessary to estimate the probability that the two profiles are a random match.

The **profile probability** or **random match probability** method gives a numerical probability that a person chosen at random from a population would share

TABLE ST 2.2 A Profile Probability Calculation Based on Analysis of Five STR Loci

STR Locus	Alleles from Profile	Allele Frequency from Population Database*	Genotype Frequency Calculation
D5S818	11	0.361	$2pq = 2 \times 0.361 \times 0.141 = 0.102$
	13	0.141	
TPOX	11	0.243	$p^2 = 0.243 \times 0.243 = 0.059$
	11	0.243	
D8S1179	13	0.305	$2pq = 2 \times 0.305 \times 0.031 = 0.019$
	16	0.031	
CSF1PO	10	0.217	$p^2 = 0.217 \times 0.217 \times 0.047$
	10	0.217	
D19S433	13	0.253	$2pq = 2 \times 0.253 \times 0.369 = 0.187$
	14	0.369	
Genotype frequency from this 5-locus profile $= 0.102 \times 0.059 \times 0.019 \times 0.047 \times 0.187 = 0.0000009 = 9 \times 10^{-7}$			

*A U.S. Caucasian population database [Butler, J. M., et al. (2003). *J. Forensic Sci*. 48:908–911].

© 2003 John Wiley & Sons, Inc.

the same DNA profile as the evidence or suspect profiles. The following example demonstrates how to arrive at a profile probability (**Table ST 2.2**).

The first locus examined in this DNA profile (D5S818) has two alleles: 11 and 13. Population studies show that the 11 allele of this locus appears at a frequency of 0.361 in this population and the 13 allele appears at a frequency of 0.141. In population genetics, the frequencies of two different alleles at a locus are given the designation *p* and q, following the Hardy—Weinberg law described earlier in the text (see Chapter 27). We assume that the person having this DNA profile received the 11 and 13 alleles at random from each parent. Therefore, the probability that this person received allele 11 from the mother and allele 13 from the father is expressed as $p \times q = pq$. In addition, the probability that the person received allele 11 from the father and allele 13 from the mother is also *pq*. Hence, the total probability that this person would have the 11, 13 genotype at this locus, by chance, is 2*pq*. As we see from Table ST 2.2, 2*pq* is 0.102 or approximately 10 percent. It is obvious from this sample that using a DNA profile of only one locus would not be very informative, as about 10 percent of the population would also have the D5S818 11, 13 genotype.

The discrimination power of the DNA profile increases when we add more loci to the analysis. The next locus of this person's DNA profile (TPOX) has two identical alleles—the 11 allele. Allele 11 appears at a frequency of 0.243 in this population. The probability of inheriting the 11 allele from each parent is $p \times p = p^2$. As we see in the table, the genotype frequency at this locus would be 0.059, which is about 6 percent of the population. If this DNA profile contained only the first two loci, we could calculate how frequently a person chosen at random from this population would have the genotype shown in the table, by multiplying the two genotype probabilities together. This would be $0.102 \times 0.059 = 0.006$. This analysis would mean that about 6 persons in 1000 (or 1 person in 166) would have this genotype. The method of multiplying all frequencies of genotypes at each locus is known as the **product rule.** It is the most frequently used method of DNA profile interpretation and is widely accepted in U.S. courts.

By multiplying all the genotype probabilities at the five loci, we arrive at the genotype frequency for this DNA profile: 9×10^{-7}. This means that approximately 9 people in every 10 million (or about 1 person in a million), chosen at random from this population, would share this 5-locus DNA profile.

The Uniqueness of DNA Profiles

As we increase the number of loci analyzed in a DNA profile, we obtain smaller probabilities of a random match. Theoretically, if a sufficient number of loci were analyzed, we could be *almost* certain that the DNA profile was unique. At the present time, law enforcement agencies in North America use a core set of 20 STR loci to generate DNA profiles. Using this 20-loci set, the probability that two people selected at random would have identical genotypes at these loci would be approximately 1×10^{-28}.

Although this would suggest that most DNA profiles generated by analysis of the 20 core STR loci would be unique on the planet, several situations can alter this interpretation. For example, identical twins share the same DNA, and their DNA profiles will be identical. Identical twins occur at a frequency of about 1 in 250 births. In addition, siblings can share one allele at any DNA locus in about 50 percent of cases and can share both alleles at a locus in about 25 percent of cases. Parents and children also share alleles, but are less likely than siblings to share both alleles at a locus. When DNA profiles come from two people who are closely related, the profile probabilities must be adjusted to take this into account. The allele frequencies and calculations that we describe here are based on assumptions that the population is large and has little relatedness or inbreeding. If a DNA profile is analyzed from a person in a small interrelated group, allele frequency tables and calculations may not apply.

DNA Profile Databases

Many countries throughout the world maintain national DNA profile databases. The first of these databases was established in the United Kingdom in 1995 and now contains more than 6 million profiles. In the United States, both state and federal governments have DNA profile databases. The entire system of databases along with tools to analyze the data is known as the **Combined DNA Index System (CODIS)** and is maintained by the FBI. As of August 2016, there were more than 16 million DNA profiles stored within the CODIS system. These include the **convicted offender database,** which contains DNA profiles from individuals convicted of certain crimes, and the **forensic database,** which contains profiles generated from crime scene evidence. In addition, some states have DNA profile databases containing profiles from suspects and from unidentified human remains and missing persons.

DNA profile databases have proven their value in many different situations. As of August 2016, use of CODIS databases had resulted in more than 350,000 profile matches that assisted criminal investigations and missing persons searches (Box 3). Despite the value of DNA profile databases, they remain a concern for many people who question the privacy and civil liberties of individuals versus the needs of the state.

ST 2.3 Technical and Ethical Issues Surrounding DNA Profiling

Although DNA profiling is sensitive, accurate, and powerful, it is important to be aware of its limitations. One limitation is that most criminal cases have either no DNA evidence for analysis or DNA evidence that would not be informative to the case. In some cases, potentially valuable DNA evidence exists but remains unprocessed and backlogged. Another serious problem is that of human error. There are cases in which innocent people have been convicted of violent crimes based on DNA samples that had been inadvertently switched during processing. DNA evidence samples from crime scenes are often mixtures derived from any number of people present at the crime scene or even from people who were not present, but whose biological material (such as hair or saliva) was indirectly introduced to the site (Box 4). Crime scene evidence

BOX 3
The Kennedy Brewer Case: Two Bite-Mark Errors and One Hit

In 1992 in Mississippi, Kennedy Brewer was arrested and charged with the rape and murder of his girlfriend's 3-year-old daughter, Christine Jackson. Although a semen sample had been obtained from Christine's body, there was not sufficient DNA for profiling. Forensic scientists were also unable to identify the ABO blood group from the bloodstains left at the crime scene. The prosecution's only evidence came from a forensic bite-mark specialist who testified that the 19 ”bite marks” found on Christine's body matched imprints made by Brewer's two top teeth. Even though the specialist had recently been discredited by the American Board of Forensic Odontology, and the defense's expert dentistry witness testified that the marks on Christine's body were actually postmortem insect bites, the court convicted Brewer of capital murder and sexual battery and sentenced him to death.

In 2001, more sensitive DNA profiling was conducted on the 1992 semen sample. The profile excluded Brewer as the donor of the semen sample. It also excluded two of Brewer's friends, and Y-chromosome profiles excluded Brewer's male relatives. Despite these test results, Brewer remained in prison for another five years, awaiting a new trial. In 2007, the Innocence Project took on Brewer's case and retested the DNA samples. The profiles matched those of another man, Justin Albert Johnson, a man with a history of sexual assaults who had been one of the original suspects in the case. Johnson subsequently confessed to Christine Jackson's murder, as well as to another rape and murder—that of a 3-year-old girl named Courtney Smith. Levon Brooks, the ex-boyfriend of Courtney's mother, had been convicted of murder in the Smith case, also based on bite-mark testimony by the same discredited expert witness.

On February 15, 2008, all charges against Kennedy Brewer were dropped, and he was exonerated of the crimes. Levon Brooks was subsequently exonerated of the Smith murder in March of 2008.

Since 1989, more than 340 people in the United States have been exonerated of serious crimes, based on DNA profile evidence. Seventeen of these people had served time on death row. In more than 140 of these exoneration cases, the true perpetrator has been identified, often through searches of DNA databases.

BOX 4

A Case of Transference: The Lukis Anderson Story

On November 30, 2012, police discovered the body of Raveesh Kumra at his home in Monte Sereno, California. Kumra's house had been ransacked, and he had suffocated from the tape used to gag him. Police collected DNA samples from the crime scene and performed DNA profiling. Several suspects were identified through matches to DNA database entries. One match, to a sample taken from Kumra's fingernails, was that of Lukis Anderson, a homeless man who was known to police. Based on the DNA profile match, Anderson was arrested, charged with murder, and jailed. He remained in jail, with a death sentence over his head, for the next five months.

The authorities believed that they had a solid case. The crime scene DNA profile was a perfect match to Anderson's DNA profile, and the lab results were accurate. Prosecutors planned to pursue the death penalty. The only problem for the prosecution was that Anderson could not have been involved in the murder, or even present at the crime scene.

On the night of the murder, Anderson had been intoxicated and barely conscious on the streets of San Jose and had been taken to the hospital, where he remained for the next 12 hours. Given his ironclad alibi, authorities were forced to release Anderson. But they remained baffled about how an innocent person's DNA could have been found on a murder victim—one whom Anderson had never even met.

Several months after Anderson's release, prosecutors announced that they had solved the puzzle. The paramedics who had treated Anderson and taken him to the hospital had then responded to the call at Kumra's house, where they had inadvertently transferred Anderson's DNA onto Kumra's fingernails. It is not clear how the transfer had occurred, but likely Anderson's DNA had been present on the paramedics' equipment or clothing.

If Lukis Anderson had not been in the hospital with an irrefutable alibi, he may have faced the death sentence based on DNA evidence. His story illustrates how too much confidence in the power of DNA evidence can lead to false accusations. It also points to the robustness of DNA, which can remain intact, survive disinfection, and be transferred from one location to another, under unlikely circumstances.

is often degraded, yielding partial DNA profiles that are difficult to interpret.

One of the most disturbing problems with DNA profiling is its potential for deliberate tampering. DNA profile technologies are so sensitive that profiles can be generated from only a few cells—or even from fragments of synthetic DNA. There have been cases in which criminals have introduced biological material to crime scenes, in an attempt to affect forensic DNA profiles. It is also possible to manufacture artificial DNA fragments that match STR loci of a person's DNA profile. In 2010, a research paper[1] reported methods for synthesizing DNA of a known STR profile, mixing the DNA with body fluids, and depositing the sample on crime scene items. When subjected to routine forensic analysis, these artificial samples generated perfect STR profiles. In the future, it may be necessary to develop methods to detect the presence of synthetic or cloned DNA in crime scene samples. It has been suggested that such detections could be done, based on the fact that natural DNA contains epigenetic markers such as methylation.

Many of the ethical questions related to DNA profiling involve the collection and storage of biological samples and DNA profiles. Such questions deal with who should have their DNA profiles stored on a database and whether police should be able to collect DNA samples without a suspect's knowledge or consent.

Another ethical question involves the use of DNA profiles that partially match those of a suspect. There are cases in which investigators search for partial matches between the suspect's DNA profile and other profiles in a DNA database. On the assumption that the two profiles arise from two genetically related individuals, law enforcement agencies pursue relatives of the person whose profile is stored in the DNA database. Testing in these cases is known as *familial DNA testing*. Should such searches be considered scientifically valid or even ethical?

As described previously, it is now possible to predict some facial features and geographical ancestries of persons based on information in their DNA sample—a method known as *DNA phenotyping*. Should this type of information be used to identify or convict a suspect?

As DNA profiling becomes more sophisticated and prevalent, we should carefully consider both the technical and ethical questions that surround this powerful new technology.

[1] Frumkin, D., et al. (2010). Authentication of forensic DNA samples. *Forensic Sci Int Genetics* 4:95—103.

Selected Readings and Resources

Journal Articles

Brettell, T. A., Butler, J. M., and Almirall, J. R. (2009). Forensic science. *Anal. Chem*. 81:4695–4711.

Butler, J. M. (2015). The future of forensic DNA analysis. *Phil. Trans. R. Soc*. B. 370:20140252. http://dx.doi.org/10.1098/rstb.2014.0252.

Enserink, M. (2011). Can this DNA sleuth help catch criminals? *Science* 331:838–840.

Frumkin, D., et al. (2010). Authentication of forensic DNA samples. *Forensic Sci. Int. Genet*. 4(2): 95–103.

Garrison, N. A.*, et al*. (2013). Forensic familial searching: Scientific and social implications. *Nat. Rev. Genet*. 14:445.

Gill, P., Jeffreys, A. J., and Werrett, D. J. (2005). Forensic applications of DNA "fingerprints." *Nature* 318:577–579.

Matheson, S. (2016). DNA Phenotyping: Snapshot of a criminal. *Cell* 166:1061–1064.

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Sci. Med. Pathol*. 5(2):77–84.

Whittall, H. (2008). The forensic use of DNA: Scientific success story, ethical minefield. *Biotechnol. J*. 3:303–305.

Zietkiewicz, E., et al. (2012). Current genetic methodologies in the identification of disaster victims and in forensic analysis. *J. Appl. Genetics* 53:41–60.

Web Sites

Brenner, C. H., Forensic mathematics of DNA matching. http://dna-view.com/profile.htm

Butler, J. M., and Reeder, D. J. Short tandem repeat DNA Internet database. http://www.cstl.nist.gov/div831/strbase/

Combined DNA Index System (CODIS), http://www.fbi.gov/about-us/lab/biometric-analysis/codis

Greenwood, V. (2016). How science is putting a new face on crime solving. *National Geographic Magazine*. http://www.nationalgeographic.com/magazine/2016/07/forensic-science-justice-crime-evidence/

Obasogie, O. K. (2013). High-tech, high-risk forensics. http://www.nytimes.com/2013/07/25/opinion/high-tech-high-risk-forensics.html?_r=0

The Innocence Project. http://www.innocenceproject.org

Review Questions

1. What is VNTR profiling, and what are the applications of this technique?
2. Why are short tandem repeats (STRs) the most commonly used loci for forensic DNA profiling?
3. Describe capillary electrophoresis. How does this technique distinguish between input DNA and amplified DNA?
4. What are the advantages and limitations of Y-chromosome STR profiling?
5. How does SNP profiling differ from STR profiling, and what are the advantages of SNP profiling?
6. Explain why mitochondrial DNA profiling is often the method of choice for identifying victims of massacres and mass disasters.
7. What is a "profile probability," and what information is required in order to calculate it?
8. Describe the database system known as CODIS. What determines whether a person's DNA profile will be entered into the CODIS system?
9. What is DNA phenotyping, and how do law enforcement agencies use this profiling method?
10. What are three major limitations of forensic DNA profiling?

Discussion Questions

1. Given the possibility that synthetic DNA could be purposely introduced into a crime scene in order into implicate an innocent person, what methods could be developed to distinguish between synthetic and natural DNA?
2. Different countries and jurisdictions have different regulations regarding the collection and storage of DNA samples and profiles. What are the regulations within your region? Do you think that these regulations sufficiently protect individual rights?
3. If you were acting as a defense lawyer in a murder case that used DNA profiling as evidence against the defendant, how would you explain to the jury the factors that might alter their interpretation of the crime scene DNA profile?
4. The phenomena of somatic mosaicism and chimerism are more prevalent than most people realize. For example, pregnancy and bone marrow transplantation may lead to a person's genome becoming a mixture of two different genomes. Describe how DNA forensic analysis may be affected by chimerism and what measures could be used to mitigate any confusion during DNA profiling. Find out more about genetic chimerism in an article by Zimmer, C., DNA double take, *New York Times,* September 16, 2013.

Genomics and Precision Medicine

Over the last decade, the terms *precision medicine* and *personalized medicine* have entered public consciousness as emerging, and likely revolutionary, approaches to disease prevention, diagnosis, and treatment. In 2015, the United States announced the Precision Medicine Initiative—a $215 million investment into the molecular tools required to bring precision medicine into routine clinical use. In the same year, the United Kingdom announced its Precision Medicine Catapult—aimed to accelerate the development and application of precision medicine technologies.

So, what are precision medicine and personalized medicine? **Precision medicine** can be defined as an individualized, molecular approach to disease diagnosis and treatment—one that examines a patient's individual genomic, proteomic, gene expression, and other molecular profiles and applies that information to select precise disease treatments and to develop new treatments and drugs. Precision medicine classifies patients into subpopulations based on their molecular profiles, and then directs each group into a treatment regimen that will bring about maximum benefit. Although often used interchangeably with precision medicine, **personalized medicine** is defined as a way to design specific, even unique, treatments for each individual, also based on their unique molecular profiles. Personalized medicine can be considered a part of precision medicine.

> "Precision medicine classifies patients into subpopulations based on their molecular profiles, and then directs each group into a treatment regimen that will bring about maximum benefit."

In this Special Topic chapter, we will examine some of the new developments in precision medicine, with an emphasis on pharmacogenomics and precision oncology. The role of gene therapy in precision medicine is discussed in Special Topic Chapter 5—Gene Therapy. Background on the genetics of cancer can be found in Chapter 24.

ST 3.1 Pharmacogenomics

Perhaps the most developed area in precision medicine is pharmacogenomics. **Pharmacogenomics** is the study of how an individual's genetic makeup determines the body's response to drugs. It also involves the development and use of drugs that are specifically targeted to a patient's genetic profile. The term *pharmacogenetics* is often used interchangeably with pharmacogenomics but refers to the study of how sequence variation within specific genes affects an individual's drug responses.

In this section, we examine two ways in which precision medicine is changing the development and use of drugs: by optimizing drug responses and by developing molecularly targeted drugs.

Optimizing Drug Responses

Every year, approximately 2 million people in the United States have serious side effects from pharmaceutical drugs and of these, approximately 100,000 will die. In addition, many patients do not respond to drug treatment as well as expected, due in part to their genetic makeup and the genomic variants that are associated with their diseases.

Sequence variations in dozens of genes affect a person's reactions to drugs. The proteins encoded by these gene variants control many aspects of drug metabolism, such as the interactions of drugs with carriers, cell-surface receptors, and transporters; with enzymes that degrade or modify drugs; and with proteins that affect a drug's storage or excretion.

Examples of genes that are involved in drug metabolism are members of the cytochrome P450 gene family. People with some cytochrome P450 gene variants metabolize and eliminate drugs slowly, which can lead to accumulations of the drug and overdose side effects. In contrast, other people have variants that cause drugs to be eliminated quickly, leading to reduced effectiveness. An example is the *CYP2D6* gene, which encodes debrisoquine hydroxylase. This enzyme is involved in the metabolism of approximately 25 percent of all pharmaceutical drugs, including acetaminophen, clozapine, beta blockers, tamoxifen, and codeine. There are more than 70 variant alleles of this gene. Some variants reduce the activity of the encoded enzyme, and others can increase it. Approximately 80 percent of people are homozygous or heterozygous for

TABLE ST 3.1 Examples of Drug Responses Affected by Gene Variants*

Gene	Drug Affected	Description
TPMT	Mercaptopurine, thioguanine, azathioprine	People with low levels of TPMT enzyme develop toxic side effects after taking thiopurine drugs for the treatment of leukemia or inflammatory conditions.
HLA-B	Allopurinol, carbamazepine, abacavir	Alleles of *HLA-B* are associated with allergic reactions to these drugs used to treat gout, epilepsy, and HIV, respectively.
CYP2D6	Codeine, tramadol, tricyclic antidepressants	Numerous alleles in the population affect metabolism of many drugs leading to underdoses and overdoses.
VKORC1	Warfarin	Warfarin anticoagulant inactivates VKORC1 protein. Variants in the *VKORC1* gene produce less protein, resulting in overdose effects at normal warfarin dosages.
CYP2C9	Warfarin	This gene encodes a liver enzyme that oxidizes warfarin. Variants metabolize warfarin less efficiently, leading to overdoses.
CYP2C19	Tricyclic antidepressants, clopidogrel, voriconazole	CYP2C19 protein is a liver enzyme that metabolizes 10—15% of drugs. Alleles result in poor metabolizers to ultra-metabolizers.
SLCO1B1	Simvastatin	This gene encodes a liver transporter. Variants are less efficient at removal of statins, which are used to control cholesterol levels.

* For more information, visit the PharmGKB Web site (https://www.pharmgkb.org/index.jsp).

the wild-type *CYP2D6* gene and are known as extensive metabolizers. Approximately 10 to 15 percent of people are homozygous for alleles that decrease activity (poor metabolizers), and the remainder of the population have duplicated genes (ultra-rapid metabolizers). Examples of other gene variants that influence the effectiveness of drugs are presented in **Table ST 3.1**.

One of the primary goals of precision medicine is to provide screening to patients prior to treatment so that the choice of drug and its dosage can be tailored to the patient's genomic profile. Normally, physicians order a single-gene test only when a specific drug needs to be prescribed or when a prescribed drug is not performing as expected. Currently, tests for genetic variants in about 20 genes are available. These tests predict reactions to approximately 100 drugs representing about 18 percent of all prescriptions in the United States. Several research hospitals have initiated programs to bring extensive genomic screening to all patients prior to treatment, and prior to development of future diseases—an approach called preemptive screening (Box 1).

Developing Targeted Drugs

Another goal of pharmacogenomics is to develop drugs that are targeted to the genetic profiles of specific subpopulations of patients. The most advanced applications are in the treatment of cancers. Large-scale sequencing studies show that each tumor is genetically unique. This genomic variability has been exploited to develop new drugs that specifically target cancer cells that may express mutant proteins or overexpress others.

One of the first success stories in precision targeted therapeutics was that of the ***HER-2*** gene and the drug **Herceptin®** in breast cancer. The *HER-2* gene codes for a transmembrane tyrosine kinase receptor protein. These receptors are located within the cell membranes of normal breast epithelial cells and, when bound to other growth factor receptors and ligands on the cell surface, they send signals to the cell nucleus that result in the transcription of genes whose products stimulate cell growth and division.

In about 25 percent of invasive breast cancers, the *HER-2* gene is amplified and the protein is overexpressed on the cell surface. The presence of *HER-2* overexpression is associated with increased tumor invasiveness, metastasis, and cell proliferation, as well as a poorer patient prognosis.

Based on this knowledge, Genentech Corporation in California developed a monoclonal antibody known as trastuzumab (or Herceptin) that binds to the extracellular region of the HER-2 receptor, inhibiting HER-2 signaling, triggering cell-cycle arrest, and leading to destruction of the cancer cell.

Because Herceptin acts only on cancer cells that have amplified *HER-2* genes, it is important to know the HER-2 status of each tumor. A number of molecular assays have been developed to determine the gene and protein status of breast cancer cells. These include immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH) assays. In IHC assays, an antibody that binds to the HER-2 protein is added to fixed tissue on a slide. The presence of bound antibody is then detected with a stain and observed under the microscope [**Figure ST 3.1(a)**]. The FISH assay (which is described in Chapter 10) assesses the number of *HER-2* genes by comparing the fluorescence signal from a HER-2 probe with a control signal from another gene that is not amplified in the cancer cells [**Figure ST 3.1(b)**].

Herceptin has had a major effect on the treatment of HER-2 positive breast cancers. When Herceptin is used in combination with chemotherapy, there is a 25 to 50 percent

BOX 1

Preemptive Pharmacogenomic Screening: The PGEN4Kids Program

Beginning in 2011, St. Jude Children's Research Hospital in Memphis, Tennessee, has offered a clinical research program entitled **PGEN4Kids.** The goal of this program is to provide patients and clinicians with pharmacogenomic screening tests for thousands of variants in hundreds of genes that may be involved in drug responses. The program is available to all incoming hospital patients, and 97 percent of patients have signed up for the program.

Using DNA derived during a single blood sample, patients are tested for approximately 2000 variants in 230 genes whose products are expected to be linked to drug responses. These tests are a combination of multigene genotyping arrays and quantitative PCR tests (described in Chapter 22). Because the effects of most of these gene variants are still unclear, the data from most of these gene tests will remain in the program's research files. However, at the present time, test results on variants in seven of the genes known to affect reactions to 23 drugs have been presented to patients and made accessible to clinicians through the patient's electronic health records. As future research reveals more associations between gene variants and drug effects, the patients' electronic health records will be automatically updated with information and alerts, to inform clinicians about how to prescribe these drugs for each individual patient. In one study, the PGEN4Kids program found that 78 percent of enrolled patients had at least one gene variant that could affect their reaction to a drug. The program also includes ongoing clinician education to introduce new information about gene–drug interactions as new research data become available.

This type of genetic screening, known as "preemptive screening" is expected to be cheaper, faster, and more efficient than ordering separate tests every time a patient is prescribed a potentially high risk drug or has had an adverse drug reaction. In 2012, a research study of patients at Vanderbilt University Medical Center revealed that almost 400 adverse drug reactions could have been avoided if clinicians had had access to a preemptive pharmacogenomic screening program.

increase in survival, compared with the use of chemotherapy alone. Herceptin has also been found effective in the treatment of other HER-2 overexpressing cancer cells, including those of the stomach and gastroesophageal junction.

There are now dozens of drugs that are targeted to the genetic status of the cancer cells (**Table ST 3.2**). For example, about 40 percent of colon cancer patients respond to the drugs **Erbitux®** (cetuximab) and **Vectibix®** (panitumumab). These two drugs are monoclonal antibodies that bind to **epidermal growth factor receptors (EGFRs)** on the surface of cells and inhibit the EGFR signal transduction pathway. To work, cancer cells must express EGFR on their surfaces and must also have a wild-type *K-ras* gene.

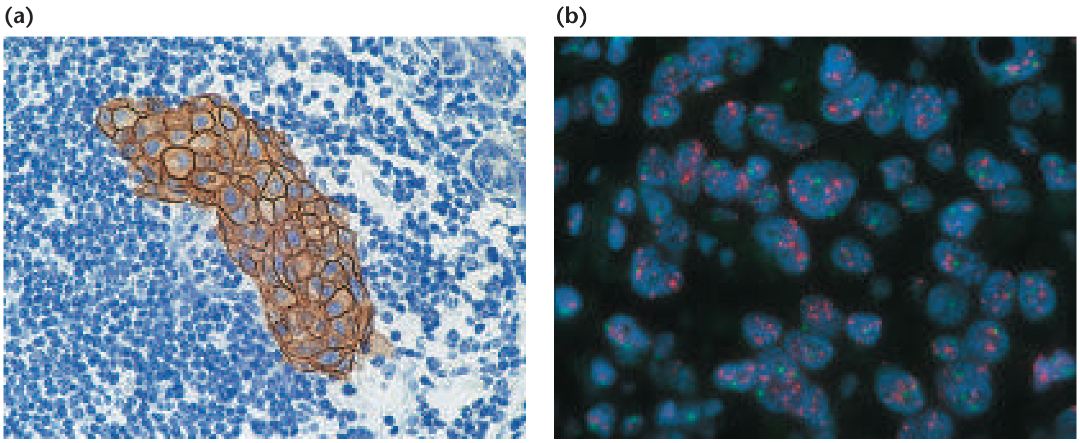


FIGURE ST 3.1 Protein and gene-amplification assays to determine HER-2 levels in cancer cells. (a) Normal and breast cancer cells within a biopsy sample, stained by HER-2 immunohistochemistry. Cell nuclei are stained blue. Cancer cells that overexpress HER-2 protein stain brown at the cell membrane. (b) Cancer cells assayed for *HER-2* gene copy number by fluorescence *in situ* hybridization. Cell nuclei are stained blue. *HER-2* gene DNA appears bright red. Chromosome 17 centromeres stain green. The degree of *HER-2* gene amplification is expressed as the ratio of red-staining foci to green-staining foci.

TABLE ST 3.2 Examples of Drugs That Specifically Target Proteins Mutated or Abnormally Expressed in Cancer Cells

Drug	Cancer Types	Target	Description
Imatinib (Gleevec)	Ph+ CML and ALL	BCR-ABL kinase	Imatinib binds to and inhibits BCR-ABL, which is encoded by the *bcl-abl* fusion gene located on the Philadelphia chromosome.
Olaparib (Lynparza)	BRCA1/2 mutated ovarian cancer	Poly ADP ribose polymerase (PARP)	BRCA1/2-defective cancers rely on PARP for DNA repair. Olaparib inhibits PARP repair, blocking cell division.
Trametinib (Mekinist)	Melanoma	Mitogen-activated protein kinase (MEK)	Trametinib inhibits mutated constitutively active MEK pathways, resulting in cell-cycle arrest and increased apoptosis.
Crizotinib (Xalkori)	NSCLC	EML4-ALK fusion kinase	Crizotinib inhibits fusion kinase activity, reducing cancer cell growth and invasion.
Vismodegib (Erivedge)	Basal-cell carcinoma	Smoothen receptor	Inhibits transcription factors that are necessary for expression of tumor genes.

Ph+ CML = Philadelphia chromosome-positive chronic myelogenous leukemia
Ph+ALL = Philadelphia chromosome-positive acute lymphoblastic leukemia
NSCLC = non-small-cell lung carcinoma

The presence of EGFR protein can be assayed using a staining test and observation of cancer cells under a microscope. Mutations in the *K-ras* gene can be detected using assays based on the polymerase chain reaction (PCR) method, which is described earlier in the text (see Chapter 20).

ST 3.2 Precision Oncology

One of the promises of precision medicine is to treat cancer patients with therapies that target specific gene mutations and gene expression defects in their tumors, leading to effective remissions and even cures. To support these promises, advances in exomic and whole-genome sequencing methods are making these technologies more cost effective for the diagnosis of many diseases including cancers. Large research programs, such as The Cancer Genome Atlas project (described in the Exploring Genomics feature in Chapter 24) are mapping the genomes of thousands of tumor types to identify mutations and expression profiles for which targeted drugs can be developed. Targeted therapies and diagnostics also benefit from high-throughput proteomic and metabolomic assays.

As described in the previous section, many cancer drugs targeted to specific genetic and gene expression profiles are already being used, sometimes with dramatic effects (Box 2). So far, the percentage of patients that can be successfully treated with precision cancer drugs is small. One clinical trial showed that only 6.4 percent of enrolled patients could be matched to a targeted drug based on their tumor's genomic profile. Another challenge is to deal with tumor resistance. To circumvent resistance, it will be necessary to use multiple treatment approaches simultaneously—both targeted and generalized.

Beyond the use of targeted drugs, researchers are also making progress in the use of other targeted modalities, including targeted cancer immunotherapies, which are described next.

Targeted Cancer Immunotherapies

Some of the most promising new developments in precision medicine are in the field of cancer **immunotherapy.** These therapies harness the patient's own immune system to kill tumors, and some have brought remarkable therapeutic effects in clinical trials and triggered billions of dollars of investment into their development. In this section, we will describe two of the most promising precision cancer immunotherapies—*adoptive cell transfer* and *engineered T-cell* methods.

To understand how these therapies work, we need to briefly review how the immune system, particularly **T cells,** defends against the development of cancer. As summarized in Box 3 and **Figure ST 3.2**, the immune system consists of cell types and chemical signals constituting the innate and the adaptive systems.

Both adoptive cell transfer and engineered T-cell methods exploit **cytotoxic T lymphocytes** (CTLs) to recognize specific antigens on the surface of cancer cells, bind to the cells, and destroy them. Box 4 and **Figure ST 3.3** summarize the steps involved in normal T-cell recognition and destruction of cancer cells.

BOX 2

Precision Cancer Diagnostics and Treatments: The Lukas Wartman Story

During his final year of medical school in 2002, Dr. Lukas Wartman began to experience symptoms of fatigue, fever, and bone pain. After months of tests, he was diagnosed with adult acute lymphoblastic leukemia (ALL). Following two years of chemotherapies, his cancer went into remission for three years. When the ALL recurred, his doctors treated him with intensive chemotherapy and a bone marrow transplant, which put him back into remission for another three years. After his second relapse, all attempts at treatment failed and he was rapidly deteriorating.

At the time of his second relapse, Dr. Wartman was working as a physician-scientist at Washington University, researching the genetics of leukemia. His colleagues, including Dr. Timothy Ley, associate director of the Washington University Genome Institute, decided to rush into a last-minute effort to save him. Using the university's sequencing facilities and supercomputers, the research team sequenced the entire genome of both his normal and his cancer cells. They also analyzed his RNA types and expression levels using RNAseq technologies.

As they had expected, Dr. Wartman's cancer cells contained many gene mutations. Unfortunately, there were no known drugs that would attack the products of these mutated genes. The RNA sequence analysis, however, revealed unexpected results.

It showed that the fms-related tyrosine kinase 3 (*FLT3*) gene, although having a normal DNA sequence, was overexpressed in his cancer cells—perhaps due to mutations in the gene's regulatory regions. The *FLT3* gene encodes a protein kinase that is involved in normal hematopoietic cell growth and differentiation, and its overexpression would be a potentially important contributor to Dr. Wartman's cancer. Equally informative, and fortunate, was that the drug sunitinib (Sutent®) was known to inhibit the FLT3 kinase and had been approved for use in the treatment of some kidney and gastrointestinal cancers.

Dr. Wartman decided to try sunitinib. Unfortunately, the drug cost $330 per day, and Dr. Wartman's insurance company refused to pay for it. In addition, the drug company Pfizer refused to supply the drug to him under its compassionate use program. Despite these setbacks, he collected enough money to buy a week's worth of sunitinib. Within days of starting treatment, his blood counts were approaching normal. Within two weeks his bone marrow was free of cancer cells. At this juncture, Pfizer reversed its decision and supplied Dr. Wartman with the drug. In addition, he underwent a second bone marrow transplant to help ensure that the cancer would not return. Although Dr. Wartman's long-term prognosis is still uncertain, his successful experience with precision cancer treatment has given him hope and has spurred research into the regulation of the *FLT3* gene in other cancers.

Cancer cells express many proteins that are specific to the tumor and have the capacity to be recognized by the patient's immune system as nonself antigens. These nonself antigens result from abnormal gene expression and mutations in the coding regions of both cancer driver and passenger genes. For example, 30 percent of human cancers contain mutated *ras*-family genes (such as *K-ras* and *H-ras*), which act as cancer driver genes. Many different point mutations can occur in these genes, each encoding an altered protein that is not found in normal cells. Cancer cells also contain up to hundreds of mutations in passenger genes whose products are not involved in the cancer phenotype, but also encode mutated, and hence nonself, proteins. Collectively, the novel, nonself antigens that are contained within their proteins are known as **neoantigens.**

Although T cells are known to associate with tumors and are able to recognize tumor neoantigens, they are often not able to destroy tumor cells. These tumor-associated T cells are also known as **tumor-infiltrating lymphocytes (TILs).**

Cancers use many different strategies to suppress T-cell responses. These strategies include the synthesis of molecules that bind to T cells and repress their activity. Interestingly, some effective new drugs called *checkpoint inhibitors* help T cells avoid these checkpoint molecules, thereby enhancing the tumor-killing ability of TILs. Another way that tumors avoid immune system activity is that they are often abnormal in their expression of cell-surface major histocompatibility complex (MHC) molecules, which are essential to stimulate antigen-presenting cells, which in turn are necessary to stimulate T cells to recognize and kill cells that bear nonself antigens (see Box 4). A third way that tumors avoid immune responses is through the presence of tumor-associated **regulatory T cells** called T-regs (including **suppressor T cells**), whose role is to repress the activities of activated T cells. The presence of other tumor-infiltrating cells such as **macrophages** and **monocytes** also repress the activities of T cells.

To circumvent and overwhelm the mechanisms that cancers use to repress anticancer immune responses, scientists have developed the following personalized T-cell—based therapies.

Adoptive cell transfer

Adoptive cell transfer (ACT) involves removing TILs from a patient's tumor, selecting those that specifically recognize tumor antigens,

BOX 3

Cell Types in the Innate and Adaptive Immune Systems

The immune system is made up of a large number of cell types and chemicals that protect the body from external and internal "nonself" entities such as bacteria, viruses, toxins, and tumor cells (Figure ST 3.2).

The **innate system** acts rapidly and nonspecifically to these agents, engulfing or degrading them. Some components also assist cells of the adaptive system.

The **adaptive (or acquired) system** destroys pathogens, tumor cells, and molecules such as toxins by recognizing and acting specifically against each entity. It does this by recognizing specific "nonself" molecules called **antigens.** Cells of the adaptive system develop a memory of previous contact with nonself antigens, allowing them to quickly replicate and respond to a subsequent appearance of the antigen.

The adaptive system has two branches. The humoral branch involves B lymphocytes (B cells) that synthesize antibodies directed at specific antigens. The cell-mediated branch consists of T lymphocytes (T cells) including cytotoxic T cells and helper T cells. These cells recognize specific antigens on the surface of or inside cells that are infected or cancerous. Cytotoxic T cells then contact the cell, release cytotoxic molecules, and trigger apoptosis of the target cell.

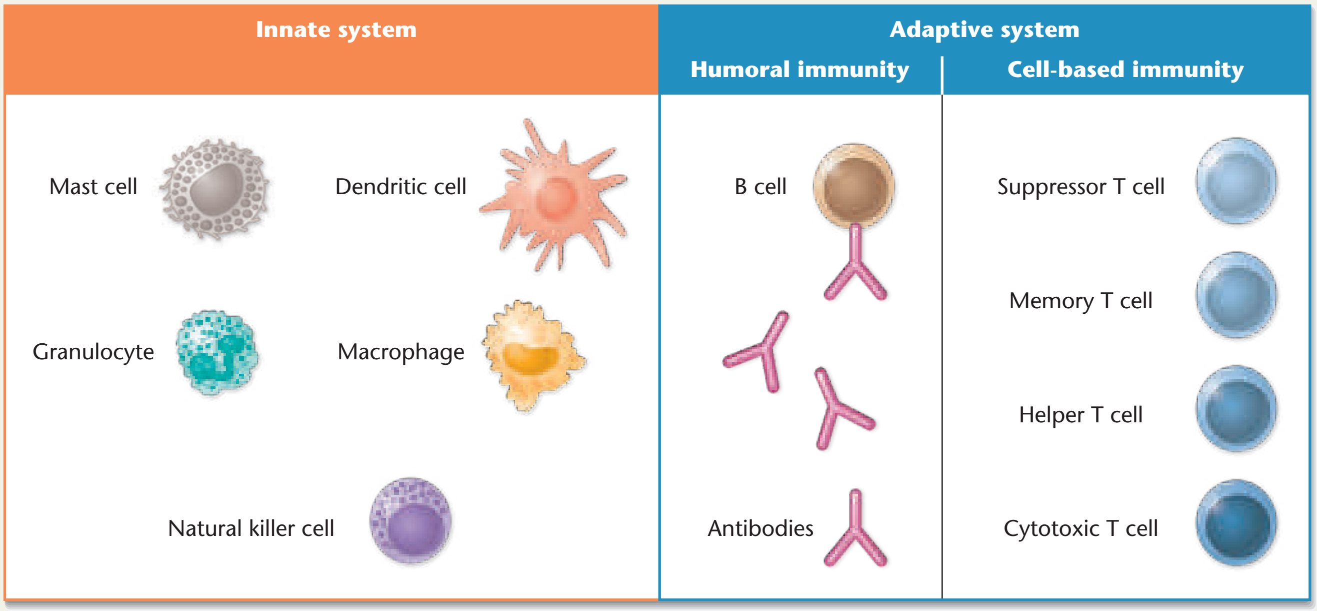


FIGURE ST 3.2 Cell types of the innate and adaptive immune systems.

amplifying these specific TILs *in vitro*, and reintroducing them back into the patient.

The steps in ACT are summarized in **Figure ST 3.4**. In the first step, tumor specimens that contain TILs are removed from the patient and digested into small samples containing one or several cells. Each sample is grown in a culture dish in the presence of tumor material and IL-2, a growth factor for T cells. As the T cells grow in the dish, those with reactivity to the tumor cells destroy the tumor cells within two to three weeks. These T cells are selected and retested for their tumor-destroying activity in coculture assays. Positive T cells are then grown to high numbers (10^{11} cells) in the lab, in the presence of several growth-stimulatory factors. The process requires about six weeks from obtaining the tumor specimen to harvesting the amplified reactive T-cell preparation. At this point, the patient is treated with chemotherapy to rid the body of immune system cells such as T-regs and macrophages that repress the activity of activated T cells. Then, the patient is reinfused with the amplified T cells and IL-2. The adoptive T cells can continue to expand up to 1000-fold after reinfusion. In some patients, the tumor-reactive T cells can be found in the circulation months after the initial infusion, where they make up as much as 80 percent of the T-cell population. The persistence of the adoptive T cells correlates with a positive antitumor effect.

The results of some ACT clinical trials have produced positive results, particularly in patients with metastatic melanoma—a cancer that normally has a poor outcome. For example, in trials conducted by the National Cancer Institute, after only one treatment, the outcome was

BOX 4

Steps in Cytotoxic T-cell Recognition, Activation, and Destruction of Cancer Cells

T cells (T lymphocytes) play a major role in cell-mediated immunity, including the recognition and destruction of cancer cells. Immature thymocytes, the precursors of T cells, originate in the bone marrow and move to the thymus where they progress through developmental stages to become naïve, inactive T cells. During their residence in the thymus, T cells undergo selection to remove those that recognize self antigens. They also begin to express specific glycoproteins on their surfaces—such as CD4 and CD8—that are characteristic of each T-cell type. Once matured, the T cells are released from the thymus into the blood and lymph nodes.

T cells express T-cell receptors (TCRs) on their cell surfaces. TCRs are transmembrane protein complexes composed of two polypeptide chains. Both of these chains have an amino-terminal variable region and a constant region, similar to antibodies. Variable regions make contact with antigens that are present on both antigen-presenting cells and target cells (infected or tumor cells). Intracellular domains of TCRs contain regions that send signals to the T cell after the variable regions make contact with a specific antigen. These signals activate the T cell, leading to T-cell proliferation and expression of gene products that give the T cells their functional capacities.

The first step in tumor-cell recognition by cytotoxic T cells (CTLs) involves the actions of antigen-presenting cells (APCs). These cells, primarily dendritic cells, scavenge proteins released from lysed tumor cells, digest these proteins into short peptides, and present the peptides on their cell surfaces, in association with cell-surface molecules known as major histocompatibility complex (MHC) molecules. MHC molecules are present on the surfaces of most cells in the body. APCs bearing MHC-antigen complexes travel to the lymph nodes where they contact naïve T cells bearing TCRs that have binding affinity for each specific antigen. After contact with the MHC antigen on an APC cell, the activated T cell proliferates and expresses “effector” gene products.

Once activated, CTLs travel through the body, until they contact a cell that expresses the antigen that is recognized by their specific TCR. This antigen, like that found on the APCs, is present on the target cell’s surface in association with an MHC molecule. CTLs bind tightly to their target cells and then release molecules such as perforin and granzymes that enter the target cells, inducing cell death through apoptosis.

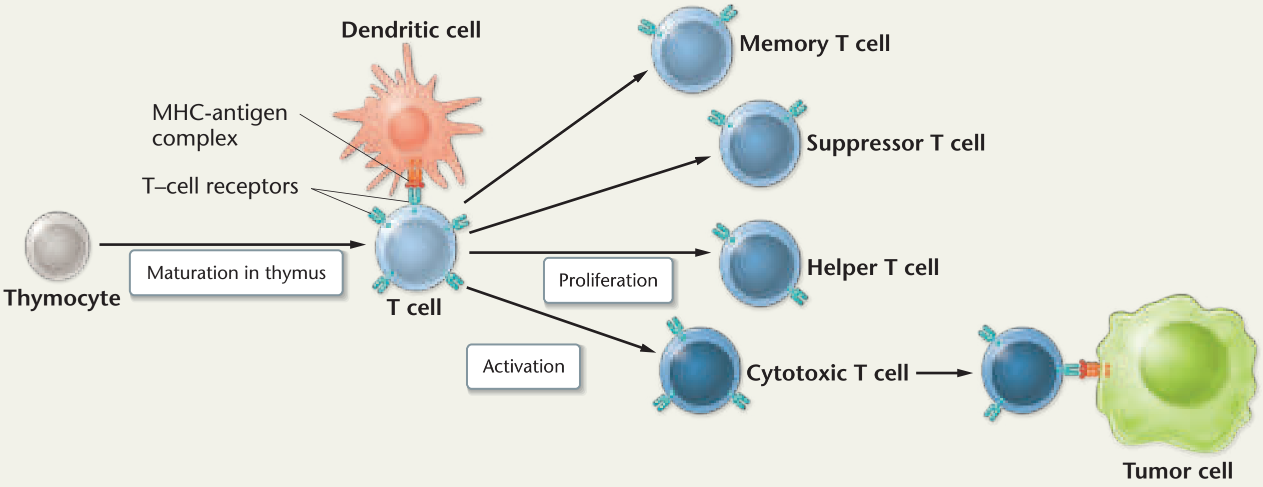


FIGURE ST 3.3 Steps in the maturation and activation of T cells.

remarkable. Between 53 and 72 percent of patients showed positive responses to the treatment, 22 percent showing complete regressions and 20 percent having no recurrence of their cancer up to 10 years later.

The promising results of ACT clinical trials on metastatic melanoma have encouraged attempts to use this method to treat other cancers. As of October 2017, researchers have reported promising clinical trials using this ACT method to treat several other cancers including cervical cancer and some blood cancers. To extend ACT therapies to patients who may not have activated TILs within their tumors that recognize unique neoantigens on those tumors, scientists are developing other ways to target the immune system to cancer. One of these methods is described next.

Immunotherapy with genetically engineered T Cells The principle behind genetically engineered T-cell therapies is to create recombinant **T-cell receptors (TCRs)** that specifically recognize antigens on cancer cells. The DNA sequences that encode these engineered TCRs are then introduced *in vitro* into a patient’s normal,

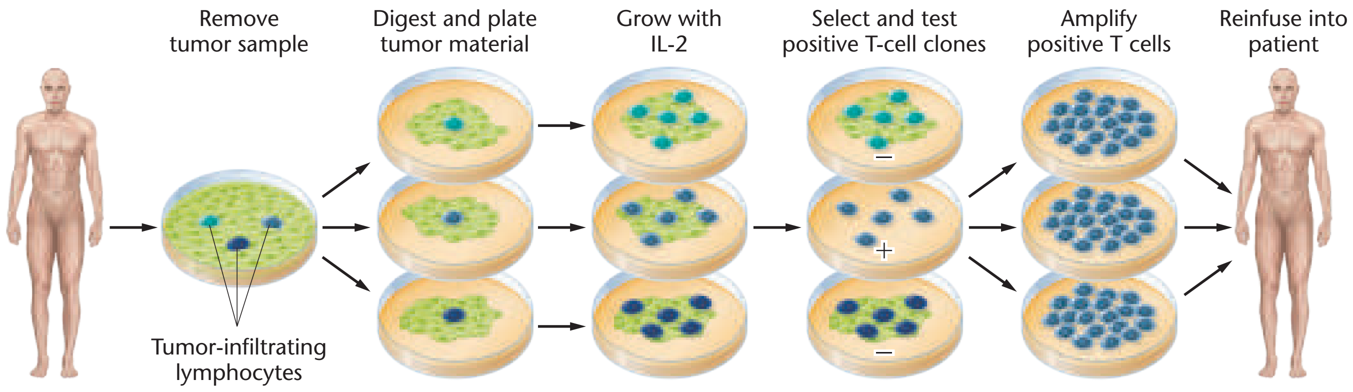

Remove tumor sample
Digest and plate tumor material
Grow with IL-2
Select and test positive T-cell clones
Amplify positive T cells
Reinfuse into patient
−
+
−
Tumor-infiltrating lymphocytes

FIGURE ST 3.4 Summary of adoptive cell transfer method.

naïve T cells which then express these TCRs on their surfaces. The TCR-transduced T cells are then selected, amplified, and reinfused into the patient. The synthetic TCR genes encode either TCRs that are structurally similar to natural TCRs or **chimeric antigen receptors (CARs)** that can directly recognize antigens on the tumor cell without requiring T-cell activation by antigen-presenting cells. In this subsection, we will describe CAR T-cell therapies.

CAR proteins are fusions of several proteins derived from a variety of sources. (The structures of normal TCRs and CARs are shown in Figure ST 3.5.) The extracellular portion of a CAR consists of the variable regions of immunoglobulin heavy and light chains, separated by a linker sequence. These variable regions fold in such a way that they mimic the specificity of an antibody that recognizes a specific antigen—such as a tumor neoantigen. The variable antibody regions are preceded by a signal peptide to direct the CAR to the surface of the T cell in which it is expressed. A spacer region allows the variable regions to orient themselves to bind to antigens on the cancer cell. A transmembrane region anchors the CAR in the T-cell membrane, and the intracellular region is responsible for sending various activation signals to the T cell, after the variable antibody regions have made contact with an antigen. The activation signals include instructions to proliferate, differentiate, produce cytokines, and kill the target cell.

To create CARs, scientists clone DNA fragments encoding each of the regions described above into a single linear recombinant DNA molecule, which encodes the entire CAR fusion protein. The DNA fragment encoding immunoglobulin variable regions is usually cloned from cells that secrete monoclonal antibodies. The cells that produce these monoclonal antibodies have previously been screened and selected for their reactivity against the desired neoantigens found on the surface of cancer cells. Once the chimeric

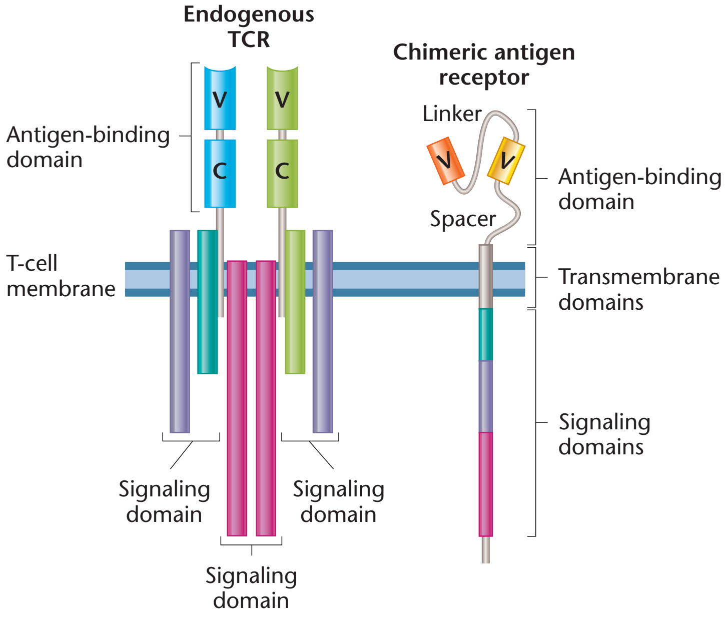


FIGURE ST 3.5 Structures of an endogenous T-cell receptor (TCR) and a recombinant chimeric antigen receptor (CAR).

DNA molecules have been cloned, they are introduced into normal T cells that have been purified from the patient's peripheral blood. The types of vectors and methods used to introduce CAR constructs into T cells are described in more detail in Special Topic Chapter 5—Gene Therapy. The remainder of the procedure is similar to adoptive cell transfer—screening and amplifying the T cells, and reinfusing them into patients who have been treated with chemotherapy or radiation therapies to reduce the numbers of endogenous immune system cells.

Results of clinical trials have been encouraging. The majority of trials have tested CAR T cells for effectiveness in treating B-cell cancers such as leukemias and lymphomas. The CAR T cells recognize the CD19 surface proteins that are expressed on B cells but not on other cells. The response rates have varied between 70 and 100 percent with reports of long-term remissions of up to several years. Several clinical trials have tested CAR T-cell therapies against solid tumors, with less encouraging results.

As of October 2017, two CAR T-cell therapies have been approved by the US Food and Drug Administration (FDA). In August 2017, a CAR T-cell therapy called Kymriah™ was approved for treating children with acute lymphoblastic leukemia (ALL) who had relapsed twice or did not respond to earlier treatments. In clinical trials of these patients, Kymriah treatment produced complete remissions in 83 percent of patients. In October 2017, the FDA approved a CAR T-cell therapy called Yescarta™ for treating adults with some types of large B-cell lymphomas.

Despite promising results with CAR T-cell therapies, these treatments have several serious side effects in many patients. These include systemic inflammatory responses, neurotoxicity, and eventual tumor resistance. Researchers expect that these side effects will become manageable as more experience is gained with these new therapies.

ST 3.3 Precision Medicine and Disease Diagnostics

The ultimate goal of personalized medicine is to apply information from a patient's full genome to help physicians diagnose disease and select treatments tailored to that particular patient. Not only will this information be gleaned from genome sequencing, but it will also be informed by gene-expression information derived from transcriptomic, proteomic, metabolomic, and epigenetic tests.

Presently, the most prevalent use of genomic information for disease diagnostics is genetic testing that examines specific disease-related genes and gene variants. Most existing genetic tests detect the presence of mutations in single genes that are known to be linked to a disease. Currently, more than 45,000 genetic tests are available. A comprehensive list of genetic tests can be viewed on the NIH Genetic Testing Registry at www.ncbi.nlm.nih.gov/gtr/. The technologies used in many of these genetic tests were presented earlier in the text (see Chapter 22).

Over the last decade, genome sequencing methods have progressed rapidly in speed, accuracy, and cost effectiveness. In addition, other "omics" technologies such as transcriptomics and proteomics are providing major insights into how DNA sequences lead to gene expression and, ultimately, to phenotype. (Refer to Chapter 21 for descriptions of techniques and data emerging from human omics technologies.) As these technologies become more rapid and cost effective, they will begin to make important contributions to precision medicine.

Although the application of omics technologies to precision diagnostics has not yet entered routine medical care, several proof-of-principle cases have illustrated the ways in which whole-genome analysis may be applied in the future. They also reveal some of the limitations that must be overcome before genome-based medicine becomes commonplace and practical.

One such case study is described in Box 5. This study combined data from whole-genome sequencing, transcriptomics, proteomics, and metabolomics profiles from a single patient at multiple time points over a 14-month period. This in-depth, multilevel personal profiling followed the patient through both healthy and diseased states, as he contracted two virus infections and a period of Type 2 diabetes. This research points out how complex changes in gene expression may affect phenotype and shows the importance of looking beyond the raw sequence of DNA. It also indicates that gene-expression profiles can be monitored by current technologies and may be applied in the future as part of personalized medical testing.

ST 3.4 Technical, Social, and Ethical Challenges

There are still many technical hurdles to overcome before precision medicine will become a standard part of medical care. The technologies of genome sequencing, omics profiling, microarray analysis, and SNP detection need to be faster, more accurate, and cheaper.

Another challenge will be the storage and interpretation of vast amounts of genomic and other omics data. For example, each personal genome generates the letter-equivalent of 200 large phone books, which must be stored in databases and mined for relevant sequence variants. Then, meaning must be assigned to each sequence variant. To undertake these kinds of analyses, scientists need to gather data from large-scale population genotyping studies that will link

BOX 5

Beyond Genomics: Personal Omics Profiling

A study published by a research team led by Dr. Michael Snyder of Stanford University provides an example of how multiple omics technologies can be used to examine one person's healthy and diseased states.[1]

Blood samples were taken from a healthy individual (Dr. Snyder) at 20 time points over a 14-month period. Dr. Snyder's whole-genome sequence was generated at each time point using two different methods and backed up by exome sequencing using three different methods. In addition, his genome sequence was compared to that of his mother. Concurrently, whole-transcriptome sequencing, proteomic profiling, and metabolomics assays were performed.

Using RNAseq technologies, the researchers monitored the numbers and types of more than 19,000 mRNAs and miRNAs transcribed from more than 12,000 genes over 20 time points. The data showed that sets of genes were coordinately regulated in response to conditions such as RSV infection and glucose levels. The researchers also found that RNA species underwent differential splicing and editing during changes in physiological states. Editing events included changes of adenosine to inosine and cytidine to uridine, and many of these RNA edits altered the amino acid sequences of translated proteins.

The researchers also profiled the levels of more than 6000 proteins and metabolites over the time course of the study. Like the RNA data, the protein and metabolite data showed coordinated changes that occurred throughout virus infections as well as glucose level changes. Some of these changes were shared between RNA, protein, and metabolites, and others were unique to each category. The medical significance of these patterns is not clear but will be addressed in future studies.

Dr. Snyder's genome sequence revealed a number of SNPs that are known to be associated with elevated risks for coronary artery disease, basal cell carcinoma, hypertriglyceridemia, and Type 2 diabetes. A mutation in the *TERT* gene, which is involved in telomere replication, is associated with an increased risk for aplastic anemia.

These omics assays were followed by a series of medical tests. Dr. Snyder had no signs of aplastic anemia, and his telomere lengths were close to normal. Similarly, his mother, who shared his mutation in the *TERT* gene, had no symptoms of aplastic anemia. Medical tests revealed he did have elevated triglyceride levels, which he subsequently controlled using medication. Blood glucose levels were initially normal but became abnormally high after he became infected with respiratory syncytial virus (RSV). In response to these data, Dr. Snyder modified his diet and exercise regime and later brought his blood glucose down to normal levels. An analysis of drug response gene variants revealed that he should have good responses to diabetes drugs, should he need them in the future.

The detailed analysis of Dr. Snyder's genome and gene expression profiles generated approximately 500,000 gigabytes of data.

Since the study was published, Dr. Snyder has expanded his research to examine another 100 people, for their genomic, proteomic, transcriptomic, and other profiles. The goal is to analyze these big-data collections to predict diseases and match profiles with targeted therapies.

[1] Chen, R. et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148:1293–1307.

sequence variants to phenotype, disease susceptibility, or drug responses. Experts suggest that such studies will take the coordinated efforts of public and private research teams more than a decade to complete. Scientists will also need to develop efficient automated systems and algorithms to deal with this massive amount of information. Precision medicine will also need to integrate information about environmental, personal lifestyle, and epigenetic factors.

Another technical challenge for precision medicine is the development of automated health information technologies. Health-care providers will need to use electronic health records to store, retrieve, and analyze each patient's genomic profile, as well as to compare this information with constantly advancing knowledge about genes and disease. Currently, approximately 10 percent of hospitals and physicians in the United States have access to these types of information technologies.

Precision medicine raises a number of societal concerns. To make precision medicine available to everyone, the costs of genetic tests, as well as the genetic counselling that accompanies them, must be reimbursed by insurance companies, even in cases where there are no prior diseases or symptoms. Regulatory changes are required to ensure that genetic tests and genomic sequencing are accurate and that the data generated are reliably stored in databases that guarantee the patient's privacy.

Precision medicine also requires changes to medical education. In the future, physicians will be expected to use genomics information as part of their patient management. For this to be possible, medical schools will need to train future physicians to interpret and explain genetic data. In addition, more genetic counsellors and genomics specialists will be required. These specialists will need to understand genomics and disease, as well as to manipulate bioinformatic data. As of 2017, there were only about 4000 licensed genetic counsellors in the United States.

The ethical aspects of precision medicine are also diverse and challenging. For example, it is sometimes argued that the costs involved in the development of genomics and precision medicine are a misallocation of limited resources. Some argue that science should solve larger problems facing humanity, such as the distribution of food and clean water, before allocating resources on precision medicine. Similarly, some critics argue that such highly specialized and expensive medical care will not be available to everyone and represents a worsening of economic inequality. There are also concerns about how we will protect the privacy of genome information that is contained in databases and private health-care records.

Most experts agree that we are at the beginning of a precision medicine revolution. Information from genetics and genomics research is already increasing the effectiveness of drugs and enabling health-care providers to predict diseases prior to their occurrence. In the future, precision and personalized medicine will touch almost every aspect of medical care. By addressing the upcoming challenges of precision medicine, we can guide its use for the maximum benefit to the greatest number of people.

Selected Readings and Resources

Journal Articles

Ashley, E. A., et al. (2010). Clinical assessment incorporating a personal genome. *Lancet* 375:1525–1535.

Hodson, R. (2016). Nature Outlook: Precision medicine. *Nature* 537:S49–S72.

McLeod, H. L. (2013). Cancer pharmacogenomics: Early promise, but concerted effort needed. *Science* 339:1563–1566.

Ormond, K. E., et al. (2010). Challenges in the clinical application of whole-genome sequencing. *Lancet* 375:1749–1751.

Pushkarev, D., et al. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotech* 27:847–850.

Rosenberg, S. A., and Restifo, N. P. (2015). Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* 348:62–68.

Ross, J. S. (2009). The HER-2 receptor and breast cancer: Ten years of targeted anti-HER-2 therapy and personalized medicine. *The Oncologist* 14:320–368.

Web Sites

Grady, D. (2016). Harnessing the immune system to fight cancer. *The New York Times*. https://www.nytimes.com/2016/07/31/health/harnessing-the-immune-system-to-fight-cancer.html.

Personalized Medicine Coalition. The case for personalized medicine. http://www.personalizedmedicinecoalition.org/Resources/The_Case_for_Personalized_Medicine

U.S. Food and Drug Administration. Table of valid genomic biomarkers in the context of approved drug labels. http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

U.S. National Institutes of Health, Genetics Home Reference. What is pharmacogenomics? http://ghr.nlm.nih.gov/handbook/genomicresearch/pharmacogenomics

Review Questions

1. What is pharmacogenomics, and how does it differ from pharmacogenetics?
2. Describe how the drug Herceptin works. What types of gene tests are ordered prior to treatment with Herceptin?
3. How do the cytochrome P450 proteins affect drug responses? Give two examples.
4. What are some of the ways that cancer cells avoid the killing effects of T cells?
5. What is adoptive cell transfer, and how is it being used to treat cancer?
6. What are chimeric antigen receptors, and how are they constructed?
7. Why is it necessary to examine gene-expression profiles, in addition to genome sequences, for effective precision medicine?
8. Using the information available on the NIH Genetic Testing Registry Web site, describe three single-gene tests that are currently used in disease diagnostics.

Discussion Questions

1. In this chapter, we present two case studies (Boxes 2 and 5) that use precision genomics analysis to predict and treat diseases. These cases have shown how precision medicine may evolve in the future and have inspired enthusiasm; however, they have also triggered concerns. What are some of these concerns, and how can they be addressed?
2. What are the biggest technical challenges that must be overcome before precision medicine becomes a routine component of medical care? What do you think is the most difficult of these challenges and why?
3. How can we ensure that a patient's privacy is maintained as genome information accumulates within medical records? How would you feel about allowing your genome sequence to be available for use in research?
4. As gene tests and genomic sequences become more commonplace, how can we prevent the emergence of "genetic discrimination" in employment and medical insurance?

Genetically Modified Foods

Throughout the ages, humans have used selective breeding techniques to create plants and animals with desirable genetic traits. By selecting organisms with naturally occurring or mutagen-induced variations and breeding them to establish the phenotype, we have evolved varieties that now feed our growing populations and support our complex civilizations.

Although we have had tremendous success shuffling genes through selective breeding, the process is a slow one. When recombinant DNA technologies emerged in the 1970s and 1980s, scientists realized that they could modify agriculturally significant organisms in a more precise and rapid way by identifying and cloning genes that confer desirable traits, then introducing these genes into organisms. Genetic engineering of animals and plants promised an exciting new phase in scientific agriculture, with increased productivity, reduced pesticide use, and enhanced flavor and nutrition.

Beginning in the 1990s, scientists created a large number of **genetically modified (GM) food** varieties. The first one, approved for sale in 1994, was the Flavr Savr tomato—a tomato that stayed firm and ripe longer than non-GM tomatoes. Soon afterward, other GM foods were developed: papaya and zucchini with resistance to virus infection, canola containing the tropical oil laurate, corn and cotton plants with resistance to insects, and soybeans and sugar beets with tolerance to agricultural herbicides. By 2012, more than 200 different GM crop varieties had been created. Worldwide, GM crops are planted on 180 million hectares of arable land, with a global value of $15 billion for GM seeds.

Although many people see great potential for GM foods—to help address malnutrition in a world with a growing human population and climate change—others question the technology, oppose GM food development, and sometimes resort to violence to stop the introduction of GM varieties (Figure ST 4.1). Even Golden Rice—a variety of rice that contains the vitamin A precursor and was developed on a humanitarian nonprofit basis to help alleviate vitamin A deficiencies in the developing world—has been the target of opposition and violence.

> "Genetic engineering of animals and plants promised an exciting new phase in scientific agriculture, with increased productivity, reduced pesticide use, and enhanced flavor and nutrition."

Some countries have outright bans on all GM foods, whereas others embrace the technologies. Opponents cite safety and environmental concerns, whereas some scientists and commercial interests extol the almost limitless virtues of GM foods. The topic of GM food attracts hyperbole and exaggerated rhetoric, information, and misinformation—on both sides of the debate.

So, what are the truths about GM foods? In this Special Topic chapter, we will introduce the science behind GM foods and examine the promises and problems of the new technologies. We will look at some of the controversies and present information to help us evaluate the complex questions that surround this topic.

ST 4.1 What Are GM Foods?

GM foods are derived from **genetically modified organisms (GMOs),** specifically plants and animals of agricultural importance. GMOs are defined as organisms whose genomes have been altered in ways that do not occur naturally. Although the definition of GMOs sometimes includes organisms that have been genetically modified by selective breeding, the most commonly used definition refers to organisms modified through genetic engineering or recombinant DNA technologies. Genetic engineering allows one or more genes to be cloned and transferred from one organism to another—either between individuals of the same species or between those of unrelated species. It also allows an organism's endogenous genes to be altered in ways that lead to enhanced or reduced expression levels. When genes are transferred between unrelated species, the resulting organism is called **transgenic.** The term **cisgenic** is sometimes used to describe gene transfers within a species. In contrast, the term **biotechnology** is a general term, encompassing a wide range of methods that manipulate organisms or their components—such as isolating enzymes or producing wine, cheese, or yogurt. Genetic modification of plants or animals is one aspect of biotechnology.

FIGURE ST 4.1 Anti-GM protesters attacking a field of genetically modified maize in southwestern France. In July 2004, hundreds of activists opposed to GM crops destroyed plants being tested by the U.S. biotech company Pioneer Hi-Bred International.

It is estimated that GM crops are grown in approximately 30 countries on 11 percent of the arable land on Earth. The majority of these GM crops (almost 90 percent) are grown in five countries—the United States, Brazil, Argentina, Canada, and India. Of these five, the United States accounts for approximately half of the acreage devoted to GM crops. According to the U.S. Department of Agriculture, 93 percent of soybeans and 90 percent of maize grown in the United States are from GM crops. In the United States, more than 70 percent of processed foods contain ingredients derived from GM crops.

Soon after the release of the Flavr Savr tomato in the 1990s, agribusinesses devoted less energy to designing GM foods to appeal directly to consumers. Instead, the market shifted toward farmers, to provide crops that increased productivity. Approximately 200 different GM crop varieties are approved for use as food or livestock feed in the United States. However, only about two dozen are widely planted. These include varieties of soybeans, corn, sugar beets, cotton, canola, papaya, and squash. **Table ST 4.1** lists some of the common GM food crops available for planting in the United States. Of these GM crops, by far the most widely planted are varieties that are herbicide tolerant or insect resistant. Only one GM food animal, the AquAdvantage salmon, has been approved for consumption (Box 1). A number of agriculturally important animals such as goats and sheep have been genetically modified to produce pharmaceutical products in their milk. The use of transgenic animals as bioreactors is discussed elsewhere in the text (see Chapter 22).

TABLE ST 4.1 Some GM Crops Approved for Food, Feed, or Cultivation in the United States*

Crop	Number of Varieties	GM Characteristics
Soybeans	19	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Reduced saturated fats Enhanced oleic acid Enhanced omega-3 fatty acid
Maize	68	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Bt insect resistance Enhanced ethanol production
Cotton	30	Tolerance to glyphosate herbicide Bt insect resistance
Potatoes	28	Bt insect resistance
Canola	23	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Enhanced lauric acid
Papaya	4	Resistance to papaya ringspot virus
Sugar beets	3	Tolerance to glyphosate herbicide
Rice	3	Tolerance to glufosinate herbicide
Zucchini squash	2	Resistance to zucchini, watermelon, and cucumber mosaic viruses
Alfalfa	2	Tolerance to glyphosate herbicide
Plum	1	Resistance to plum pox virus

*Information from the International Service for the Acquisition of Agri-Biotech Applications, www.isaaa.org.

Herbicide-Resistant GM Crops

Weed infestations destroy about 10 percent of crops worldwide. To combat weeds, farmers often apply herbicides before seeding a crop and between rows after the crops are growing. As the most efficient broad-spectrum herbicides also kill crop plants, herbicide use may be difficult and limited. Farmers also use tillage to control weeds; however, tillage damages soil structure and increases erosion.

Herbicide-tolerant varieties are the most widely planted of GM crops, making up approximately 70 percent of all GM crops. The majority of these varieties contain a bacterial gene that confers tolerance to the broad-spectrum herbicide **glyphosate**—the active ingredient in commercial herbicides such as Roundup®. Studies have shown that glyphosate is effective at low concentrations, is degraded rapidly in soil and water, and is not toxic to humans.

Farmers who plant glyphosate-tolerant crops can treat their fields with glyphosate, even while the GM crop is growing. This approach is more efficient and economical than mechanical weeding and reduces soil damage caused by repeated tillage. It is suggested that there is less environmental impact when using glyphosate, compared with having to apply higher levels of other, more toxic, herbicides.

Recently, evidence suggests that some weeds may be developing resistance to glyphosate, thereby reducing the effectiveness of glyphosate-tolerant crops. (This and other concerns about herbicide-tolerant GM plants are described later in this chapter.) One method used to engineer a glyphosate-tolerant plant is described in the next section.

BOX 1

The Tale of GM Salmon—Downstream Effects?

In 2015, the AquAdvantage salmon became the first GM animal to be approved for human consumption.

The AquAdvantage salmon is an Atlantic salmon that is genetically modified to grow twice as fast as its non-GM cousins, reaching marketable size in one and a half years rather than the usual three years. Scientists at AquaBounty Technologies in Massachusetts created the variety by transforming an Atlantic salmon with a single gene encoding the Chinook salmon growth hormone. The gene was cloned downstream of the antifreeze protein gene promoter from an eel. This promoter stimulates growth hormone synthesis in the winter, a time when the fish's own growth hormone gene is not expressed. The rapid growth of the GM salmon allows fish farmers to double their productivity.

AquaBounty will produce the salmon eggs at a facility in Canada and grow the salmon in a containment facility in Panama. To ensure that the fish will not escape the facilities, the company promises to sell only fertilized eggs that are female, triploid, and sterile. The facilities have tanks that are located inland and have sufficient filters to ensure that eggs and small fish cannot escape.

Despite these assurances, environmental groups are planning to fight the sale of GM salmon. Some grocery chains in the United States have banned GM fish.

Supporters of GM fish point out that the GM salmon are very unlikely to escape their facilities, and if any did escape, they would be poorly adapted to wild conditions. Critics of the new GM salmon point out that the technique used to create sterile triploids (pressure-shocking the fertilized eggs) still allows a small percentage of fertile diploids to remain in the stock. They state that even a few fertile fish, if they escaped into the wild, could have long-term effects on wild populations. A study published in 2013 shows that it is possible for the AquAdvantage salmon to breed successfully with a close relative, the brown trout.* In laboratory conditions, the hybrids grew more quickly than either the GM or non-GM varieties, and in closed stream-like systems, the hybrids outcompeted both parental fish varieties for food supplies. It is still not known whether the hybrid salmon–trout variety could successfully breed in the wild. If GM salmon could escape, breed, and introduce transgenes into wild populations, there could be unknown negative downstream effects on fish ecosystems.

* Oke, K. B., et al. (2013). Hybridization between genetically modified Atlantic salmon and wild brown trout reveals novel ecological interactions. *Proc. R. Soc. B.* 280 (1763): 20131047.

The AquAdvantage salmon grows twice as fast as a non-GM Atlantic salmon, reaching market size in half the time.

Insect-Resistant GM Crops

The second most prevalent GM modifications are those that make plants resistant to agricultural pests. Insect damage is one of the most serious threats to worldwide food production. Farmers combat insect pests using crop rotation and predatory organisms, as well as applying insecticides.

The most widely used GM insect-resistant crops are the **Bt crops.** ***Bacillus thuringiensis*** (Bt) is group of soil-dwelling bacterial strains that produce crystal (Cry) proteins that are toxic to certain species of insects. These Cry proteins are encoded by the bacterial *cry* genes and form crystal structures during sporulation. The Cry proteins are toxic to Lepidoptera (moths and butterflies), Diptera (mosquitoes and flies), Coleoptera (beetles), and Hymenoptera (wasps and ants). Insects must ingest the bacterial spores or Cry proteins in order for the toxins to act. Within the high pH of the insect gut, the crystals dissolve and are cleaved by insect protease enzymes. The Cry proteins bind to receptors on the gut wall, leading to breakdown of the gut membranes and death of the insect.

Each insect species has specific types of gut receptors that will match only a few types of Bt Cry toxins. As there are more than 200 different Cry proteins, it is possible to select a Bt strain that will be specific to one pest type.

Bt spores have been used for decades as insecticides in both conventional and organic gardening, usually applied in liquid sprays. Sunlight and soil rapidly break down the Bt insecticides, which have not shown any adverse effects on groundwater, mammals, fish, or birds. Toxicity tests on humans and animals have shown that Bt causes few negative effects.

To create Bt crops, scientists introduce one or more cloned *cry* genes into plant cells using methods described in the next section. The GM crop plants will then manufacture their own Bt Cry proteins, which will kill the target pest species when it eats the plant tissues.

Although Bt crops have been successful in reducing crop damage, increasing yields, and reducing the amounts of insecticidal sprays used in agriculture, they are also controversial. Early studies suggested that Bt crops harmed Monarch butterfly populations, although more recent studies have drawn opposite conclusions. Other concerns still exist, and these will be discussed in subsequent sections of this chapter.

GM Crops for Direct Consumption

To date, most GM crops have been designed to help farmers increase yields. Also, most GM food crops are not consumed directly by humans, but are used as animal feed or as sources of processed food ingredients such as oils, starches, syrups, and sugars. For example, 98 percent of the U.S. soybean crop is used as livestock feed. The remainder is processed into a variety of food ingredients, such as lecithin, textured soy proteins, soybean oil, and soy flours. However, a few GM foods have been developed for direct consumption. Examples are rice, squash, and papaya.

One of the most famous and controversial examples of GM foods is **Golden Rice**—a rice variety designed to synthesize beta-carotene (the precursor to **vitamin A**) in the rice grain endosperm.

Vitamin A deficiency is a serious health problem in more than 60 countries, particularly countries in Asia and Africa. The World Health Organization estimates that 190 million children and 19 million pregnant women are vitamin A deficient. Between 250,000 and 500,000 children with vitamin A deficiencies become blind each year, and half of these will die within a year of losing their sight. As vitamin A is also necessary for immune system function, deficiencies lead to increases in many other conditions, including diarrhea and virus infections. The most seriously affected people live in the poorest countries and have a basic starch-centered diet, often mainly rice. Vitamin A is normally found in dairy products and can be synthesized in the body from beta-carotene found in orange-colored fruits and vegetables and in green leafy vegetables.

Several approaches are being taken to alleviate the vitamin A deficiency status of people in developing countries. These include supplying high-dose vitamin A supplements and growing fresh fruits and vegetables in home gardens. These initiatives have had partial success, but the expense of delivering education and supplementation has impeded the effectiveness of these programs.

In the 1990s, scientists began to apply recombinant DNA technology to help solve vitamin A deficiencies in people with rice-based diets. Although the rice plant naturally produces beta-carotene in its leaves, it does not produce it in the rice grain endosperm, which is the edible part of the rice. The beta-carotene precursor, geranylgeranyl-diphosphate, is present in the endosperm, but the enzymes that convert it to beta-carotene are not synthesized (**Figure ST 4.2**).

In the first version of Golden Rice, scientists introduced the genes *phytoene synthase* (*psy*) cloned from the daffodil plant and *carotene desaturase* (*crtI*) cloned from the bacterium *Erwinia uredovora* into rice plants. The bacterial *crtI* gene was chosen because the enzyme encoded by this gene can perform the functions of two of the missing rice enzymes, thereby simplifying the transformation process. The resulting plant produced rice grains that were a yellow color due to the presence of beta-carotene (**Figure ST 4.3**). This strain synthesized modest levels of beta-carotene—but only enough to potentially supply 15–20 percent of the recommended daily allowance of vitamin A. In the second version of the GM plant, called Golden Rice 2, the daffodil *psy* gene was replaced with the *psy* gene from maize. Golden Rice 2 produced beta-carotene levels that were more than

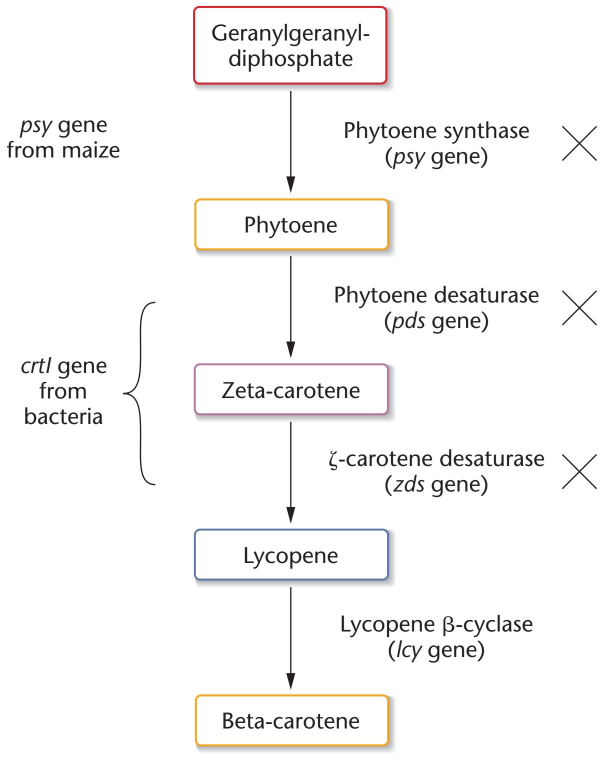

Geranylgeranyl-diphosphate
psy gene from maize
Phytoene synthase (*psy* gene)
Phytoene
Phytoene desaturase (*pds* gene)
crtI gene from bacteria
Zeta-carotene
ζ-carotene desaturase (*zds* gene)
Lycopene
Lycopene β-cyclase (*lcy* gene)
Beta-carotene

FIGURE ST 4.2 Beta-carotene pathway in Golden Rice 2. Rice plant enzymes and genes involved in beta-carotene synthesis are shown on the right. The enzymes that are not expressed in rice endosperm are indicated with an "X." The genes inserted into Golden Rice 2 are shown on the left.

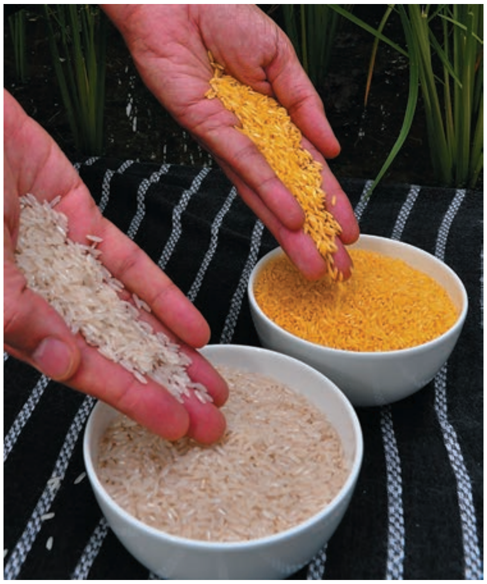


FIGURE ST 4.3 Non-GM and Golden Rice 2. Golden Rice 2 contains high levels of beta-carotene, giving the rice endosperm a yellow color. The intensity of the color reflects the amount of beta-carotene in the endosperm.

20-fold greater than those in Golden Rice. In the next section we describe the methods used to create Golden Rice 2.

Clinical trials have shown that the beta-carotene in Golden Rice 2 is efficiently converted into vitamin A in humans and that about 150 grams of uncooked Golden Rice 2 (which is close to the normal daily rice consumption of children aged 4–8 years) would supply all of the childhood daily requirement for vitamin A.

At the present time, Golden Rice 2 is undergoing field, biosafety, and efficacy testing in preparation for approval by government regulators in Bangladesh and the Philippines. If Golden Rice 2 proves useful in alleviating vitamin A deficiencies and is approved for use, seed will be made available at the same price as non-GM seed and farmers will be allowed to keep and replant seed from their own crops.

Despite the promise of Golden Rice 2, controversies remain. Critics of GM foods suggest that Golden Rice could make farmers too dependent on one type of food or might have long-term health or environmental effects. These and other controversies surrounding GM foods are discussed in subsequent sections of this chapter.

ST 4.2 Methods Used to Create GM Plants

Most GM plants have been created using one of two approaches: the **biolistic method** or ***Agrobacterium tumefaciens*–mediated transformation** technology. Both methods target plant cells that are growing *in vitro*. Scientists can generate plant tissue cultures from various types of plant tissues, and these cultured cells will grow either in liquid cultures or on the surface of solid growth media. When grown in the presence of specific nutrients and hormones, these cultured cells will form clumps of cells called calluses, which, when transferred to other types of media, will form roots. When the rooted plantlets are mature, they are transferred to soil medium in greenhouses where they develop into normal plants.

The *biolistic method* is a physical method of introducing DNA into cells. Particles of heavy metals such as gold are coated with the DNA that will transform the cells; these particles are then fired at high speed into plant cells *in vitro,* using a device called a **gene gun.** Cells that survive the bombardment may take up the DNA-coated particles, and the DNA may migrate into the cell nucleus and integrate into a plant chromosome. Plants that grow from the bombarded cells are then selected for the desired phenotype.

Although biolistic methods are successful for a wide range of plant types, a much improved transformation rate is achieved using *Agrobacterium-mediated technology*. *Agrobacterium tumefaciens* (also called *Rhizobium radiobacter*) is a soil microbe that can infect plant cells and cause tumors. These characteristics are conferred by a 200-kb tumor-inducing plasmid called a **Ti plasmid.** After infection with *Agrobacterium*, the Ti plasmid integrates a segment of its DNA known as transfer DNA (T-DNA) into random locations within the plant genome (**Figure ST 4.4**). To use the Ti plasmid as a transformation vector, scientists remove the T-DNA segment and replace it with cloned DNA of the genes to be introduced into the plant cells.

In order to have the newly introduced gene expressed in the plant, the gene must be cloned next to an appropriate promoter sequence that will direct transcription in the required plant tissue. For example, the beta-carotene pathway genes introduced into Golden Rice were cloned next to a promoter that directs transcription of the genes in the rice endosperm. In addition, the transformed gene requires appropriate transcription termination signals and signal sequences that allow insertion of the encoded protein into the correct cell compartment.

Selectable Markers

The rates at which T-DNA successfully integrates into the plant genome and becomes appropriately expressed are low. Often, only one cell in 1000 or more will be successfully transformed. Before growing cultured plant cells into mature plants to test their phenotypes, it is important to eliminate the background of nontransformed cells. This can be done using either positive or negative selection techniques.

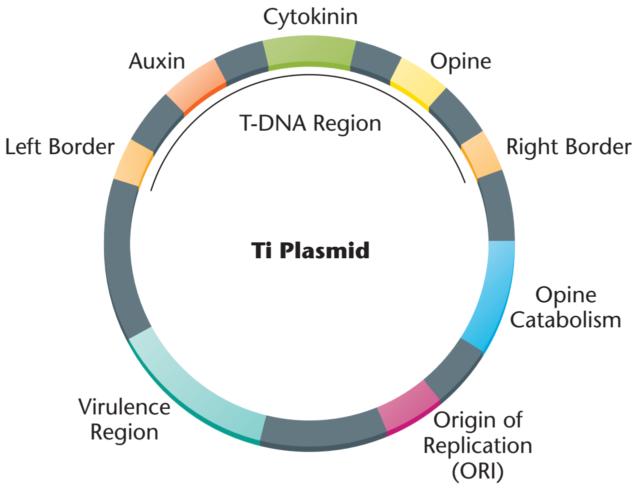


FIGURE ST 4.4 Structure of the Ti plasmid. The 250-kb Ti plasmid from *Agrobacterium tumefaciens* inserts the T-DNA portion of the plasmid into the host cell's nuclear genome and induces tumors. Genes within the virulence region code for enzymes responsible for transfer of T-DNA into the plant genome. The T-DNA region contains auxin and cytokinin genes that encode hormones responsible for cell growth and tumor formation. The opine genes encode compounds used as energy sources for the bacterium. The T-DNA region of the Ti plasmid is replaced with the gene of interest when the plasmid is used as a transformation vector.

An example of negative selection involves use of a **marker gene** such as the hygromycin-resistance gene. This gene, together with an appropriate promoter, can be introduced into plant cells along with the gene of interest. The cells are then incubated in culture medium containing hygromycin—an antibiotic that also inhibits the growth of eukaryotic cells. Only cells that express the hygromycin-resistance gene will survive. It is then necessary to verify that the resistant cells also express the cotransformed gene. This is often done by techniques such as PCR amplification using gene-specific primers. Plants that express the gene of interest are then tested for other characteristics, including the phenotype conferred by the introduced gene of interest.

An example of positive selection involves the use of a selectable marker gene such as that encoding **phosphomannose isomerase (PMI).** This enzyme is common in animals but is not found in most plants. It catalyzes the interconversion of mannose 6-phosphate and fructose 6-phosphate. Plant cells that express the *pmi* gene can survive on synthetic culture medium that contains only mannose as a carbon source. Cells that are cotransformed with the *pmi* gene under control of an appropriate promoter and the gene of interest can be positively selected by growing the plant cells on a mannose-containing medium. This type of positive selection was used to create Golden Rice 2. Studies have shown that purified PMI protein is easily digested, nonallergenic, and nontoxic in mouse oral toxicity tests. A variation in positive selection involves use of a marker gene whose expression results in a visible phenotype, such as deposition of a colored pigment.

The following descriptions illustrate the methods used to engineer two GM crops: Roundup-Ready soybeans and Golden Rice 2.

Roundup-Ready® Soybeans

The Roundup-Ready soybean GM variety received market approval in the United States in 1996. It is a GM plant with resistance to the herbicide glyphosate, the active ingredient in Roundup, a commercially available broad-spectrum herbicide. Glyphosate interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), which is present in all plants and is necessary for plant synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan. EPSPS is not present in mammals, which obtain aromatic amino acids from their diets.

To produce a glyphosate-resistant soybean plant, researchers cloned an *epsps* gene from the *Agrobacterium* strain CP4. This gene encodes an EPSPS enzyme that is resistant to glyphosate. They then cloned the *CP4 epsps* gene downstream of a constitutively expressed promoter from the cauliflower mosaic virus to allow gene expression in all plant tissues. In addition, a short peptide known as a chloroplast transit peptide (in this case from petunias) was cloned onto the $5'$-end of the *epsps* gene-coding sequence. This allowed newly synthesized EPSPS protein to be inserted into the soybean chloroplast (**Figure ST 4.5**). The final plasmid contained two *CP4 epsps* genes and, for the initial experiments, a *beta-glucuronidase* (*GUS*) gene from *E. coli*. The *GUS* gene acted as a positive marker, as cells that expressed the plasmid after transformation could be detected by the presence of a blue precipitate. The final cell line chosen for production of Roundup-Ready soybeans did not contain the *GUS* gene.

The plasmids were introduced into cultured soybean cells using biolistic bombardment. Afterward, cells were treated with glyphosate to eliminate any nontransformed cells (**Figure ST 4.6**). The resulting calluses were grown

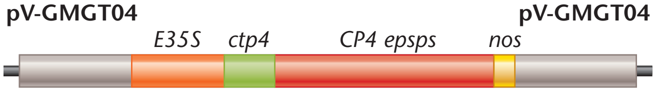


FIGURE ST 4.5 Portion of plasmid pV-GMGT04 used to create Roundup-Ready soybeans. A 1365-bp fragment encoding the EPSPS enzyme from *Agrobacterium CP4* was cloned downstream from the cauliflower mosaic virus *E35S* promoter and the petunia chloroplast transit peptide signal sequence (*ctp4*). CTP4 signal sequences direct the EPSPS protein into chloroplasts, where aromatic amino acids are synthesized. The *CP4 epsps* coding region was cloned upstream of the nopaline synthase (*nos*) transcription termination and polyadenylation sequences. The *CP4 epsps* sequences encode a 455-amino-acid 46-kDa ESPSP protein.

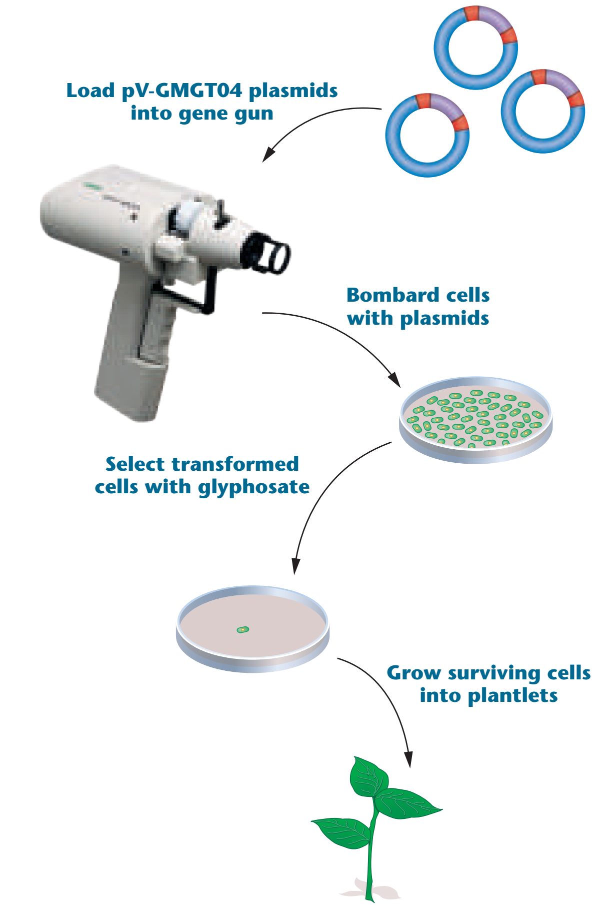

Load pV-GMGT04 plasmids into gene gun
Bombard cells with plasmids
Select transformed cells with glyphosate
Grow surviving cells into plantlets

FIGURE ST 4.6 Method for creating Roundup-Ready soybeans. Plasmids were loaded into the gene gun and fired at high pressure into cells growing in tissue cultures. Cells were grown in the presence of glyphosate to select those that had integrated and expressed the *epsps* gene. Surviving cells were stimulated to form calluses and to grow into plantlets.
© 2018 Courtesy of Bio-Rad Laboratories, Inc.

into plants, which were then field tested for glyphosate resistance and a large number of other parameters, including composition, toxicity, and allergenicity.

Golden Rice 2

To create Golden Rice 2, scientists cloned three genes into the T-DNA region of a Ti plasmid. The Ti plasmid, called pSYN12424, is shown in Figure ST 4.7. The first gene was the *carotene desaturase* (*crtI*) gene from *Erwinia uredovora,* fused between the rice *glutelin* gene promoter (*Glu*) and the *nos* gene terminator region (*nos*). The *Glu* promoter directs transcription of the fusion gene specifically in the rice endosperm. The *nos* terminator was cloned from the *Agrobacterium tumefaciens nopaline synthase* gene and supplies the transcription termination and polyadenylation sequences required at the $3'$-end of plant genes. The second gene was the *phytoene synthase* (*psy*) gene cloned from maize. The maize *psy* gene has approximately 90 percent sequence similarity to the rice *psy* gene and is involved in carotenoid synthesis in maize endosperm. This gene was also fused to the *Glu* promoter and the *nos* terminator sequences in order to obtain proper transcription initiation and termination in rice endosperm. The third gene was the selectable marker gene *phosphomannose isomerase (pmi*), cloned from *E. coli*. In the Golden Rice 2 Ti plasmid, the *pmi* gene was fused to the maize *polyubiquitin* gene promoter (*Ubi1*) and the *nos* terminator sequences. The *Ubi1* promoter is a constitutive promoter, directing transcription of the *pmi* gene in all plant tissues.

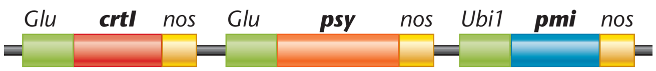


FIGURE ST 4.7 T-DNA region of Ti plasmid pSYN12424. The Ti plasmid used to create Golden Rice 2 contained the *carotene desaturase* (*crtI*) gene cloned from bacteria, the *phytoene synthase* (*psy*) gene cloned from maize, and the *phosphomannose isomerase* (*pmi*) gene cloned from *E. coli*. The *glutelin* (*Glu*) gene promoter directs transcription in rice endosperm, and the *polyubiquitin* (*Ubi1*) promoter directs transcription in all tissues. Transcription termination signals were provided by the *nopaline synthase* (*nos*) gene $3'$ region.

To introduce the pSYN12424 plasmid into rice cells, researchers established embryonic rice cell cultures and infected them with *Agrobacterium tumefaciens* that contained pSYN12424 (Figure ST 4.8). The cells were then placed under selection, using culture medium containing only mannose as a carbon source. Surviving cells expressing the *pmi* gene were then stimulated to form calluses that were grown into plants. To confirm that all three genes were present in the transformed rice plants, samples were taken and analyzed by the polymerase chain reaction (PCR) using gene-specific primers. Plants that contained one integrated copy of the transgenic construct and synthesized beta-carotene in their seeds were selected for further testing.

Gene Editing and GM Foods

The previously described methods are those that have been used to create the majority of GM plants. In the last few years, several new and revolutionary methods of genome modification have entered the field of GM foods. Collectively, these are known as **gene-editing methods.** They include zinc-finger nuclease (ZFN), transcription activator-like effector nuclease (TALEN), and CRISPR-Cas techniques. These methods are described in detail in Special Topic Chapter 1—CRISPR-Cas and Genome Editing and Special Topic Chapter 5—Gene Therapy.

Gene-editing methods have had significant effects on the speed at which scientists can induce genetic changes in plants and animals as well as on the types of changes that are possible. Gene editing allows researchers to create precise nucleotide or single-gene mutations or deletions without introducing foreign DNA into the organism. To create gene-edited plants or animals, scientists typically mutate or inactivate only one or two of the organism's endogenous genes.

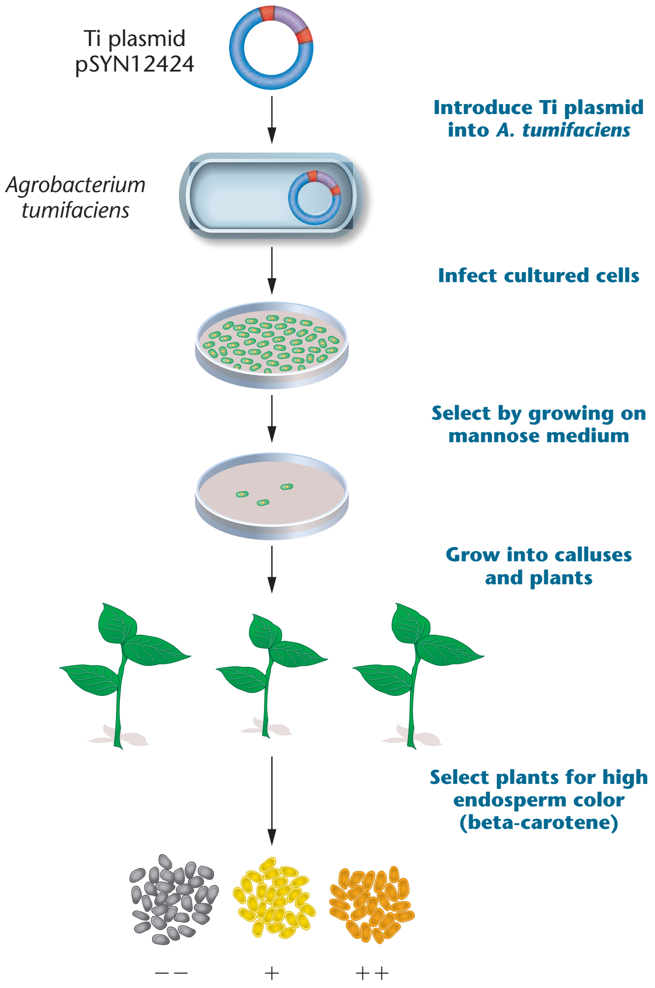

Ti plasmid pSYN12424
Introduce Ti plasmid into *A. tumifaciens*
Agrobacterium tumifaciens
Infect cultured cells
Select by growing on mannose medium
Grow into calluses and plants
Select plants for high endosperm color (beta-carotene)
− −
+
+ +

FIGURE ST 4.8 Method for creating Golden Rice 2. Rice plant cells were transformed by pSYN12424 and selected on mannose-containing medium, as described in the text. Plants that produced high levels of beta-carotene in rice grain endosperm (+ +), based on the intensity of the grain's yellow color, were selected for further analysis.

Because gene-edited organisms contain no transgene material, they have not been considered genetically modified by most regulatory agencies and therefore do not have the same oversight as other GM foods. As of January 2017, hundreds of acres of gene-edited crops have been planted in the United States, and some have been sold for human consumption (Box 2).

An example of a gene-edited food is a potato developed by the biotechnology company Calyxt, Inc. Using TALEN methods, they inactivated the *vacuolar invertase* gene that encodes an enzyme responsible for degrading sugars in cold-stored potatoes. This gene inactivation resulted in a potato with an increased storage life as well as one that does not produce harmful acrylamides when the potato is fried. The gene-edited potato was approved by the U.S. Department of Agriculture in 2014 and is now in field trials.

Scientists are also using gene-editing technologies to introduce gene alterations into farm animals. For example, the Roslin Institute in Scotland is developing a strain of pigs that is immune to the African swine fever virus. Using ZFN and CRISPR-Cas9 methods, they have introduced small changes to one of the pigs' immune system genes (*RELA*) so that the gene has the same DNA sequence as the *RELA* gene from warthogs which are resistant to the virus. The pigs are now in infection trials. If successful, the pigs could help farmers in sub-Saharan Africa and Eastern Europe where the disease is endemic.

Another example is that of the double-muscled pig, developed at Seoul National University. Using TALEN methods, researchers introduced mutations into both copies of the myostatin (*MSTN*) gene, inactivating it. Myostatin is responsible for inhibiting the growth of muscle cells. When the gene is inactivated, muscle tissue grows to produce muscle-enhanced animals. Such animals produce higher yields of lean meat.

Although gene-edited crops and animals are currently not regulated in the same way as GM foods, some countries are reviewing their guidelines and new regulations may be introduced in the near future.

ST 4.3 GM Foods Controversies

GM foods may be the most contentious of all products of modern biotechnology. Advocates of GM foods state that the technologies have increased farm productivity, reduced pesticide use, preserved soils, and have the potential to feed growing human populations. Critics claim that GM foods are unsafe for both humans and the environment; accordingly, they are applying pressure on regulatory agencies to ban or severely limit the extent of GM food use. These campaigns have affected regulators and politicians, resulting in a patchwork of regulations throughout the world. Often the debates surrounding GM foods are highly polarized and emotional, with both sides in the debate exaggerating their points of view and selectively presenting the data. So, what are the truths behind these controversies?

One point that is important to make as we try to answer this question is that *it is not possible to make general statements about all "GM foods."* Each GM crop or organism contains different genes from different sources, attached to different expression sequences, accompanied by different marker or selection genes, inserted into the genome in different ways and in different locations. In addition, the recent proliferation of gene-edited food organisms further complicates the situation. Many advocates and regulatory agencies state that gene-edited foods are not GM foods, as they do not fit the previous definitions of GMOs. GM foods are created for different purposes and are used in ways that are both planned and unplanned. Each construction is unique and therefore needs to be assessed separately.

BOX 2

The New CRISPR Mushroom

The common white-button mushroom (*Agaricus bisporus*) has become the first gene-edited crop, created using the *CRISPR-Cas9* method, to be approved for commercialization and human consumption. In April 2016, the U.S. Department of Agriculture (USDA) declared that the new mushroom, gene-edited to reduce browning, would not be subject to its regulation. Approximately 30 other GM foods, created using other gene-editing tools such as *ZNF* and *TALENS*, have gained USDA approval over the last five years.

The nonbrowning mushroom was developed by Dr. Yinong Yang of Pennsylvania State University. Dr. Yang's team used the CRISPR-Cas9 gene-editing method (described in Special Topic Chapter 1–CRISPR-Cas and Genome Editing) to inactivate one member of the mushroom's *polyphenol oxidase (PPO)* gene family, resulting in a 30 percent reduction in the levels of PPO enzyme activity. The PPO enzyme is responsible for the browning effect when mushrooms are bruised or cut. The new mushroom has a longer shelf life and resists browning caused by processing and mechanical harvesting.

Although not legally required to do so, Dr. Yang plans to voluntarily submit the mushroom to the U.S. Food and Drug Administration for safety clearance prior to commercialization.

The CRISPR mushroom was approved by the USDA because it was not created using traditional GM methods that involve introducing foreign DNA—from plasmids, viruses, or genes from other species—into the organism. In addition, the genetic changes were small (a few nucleotides) and precise, at known locations in the genome. Proponents of GM foods see the success of the new gene-edited foods as a boost for GM food development. It is still not clear how the public or GM food opponents will accept the CRISPR mushroom or other new gene-edited foods.

The CRISPR-edited mushroom resists the browning that occurs when mushrooms are cut or bruised during processing or storage, as seen here.

We will now examine two of the main GM foods controversies: those involving human health and safety, and environmental effects.

Health and Safety

GM food advocates often state that there is no evidence that GM foods currently on the market have any adverse health effects, either from the presence of toxins or from potential allergens. These conclusions are based on two observations. First, humans have consumed several types of GM foods for more than 20 years, and no reliable reports of adverse effects have emerged. Second, the vast majority of toxicity tests in animals, which are required by government regulators prior to approval, have shown no negative effects. A few negative studies have been published, but these have been criticized as poorly executed or nonreproducible.

Critics of GM foods counter the first observation in several ways. First, as described previously, few GM foods are eaten directly by consumers. Instead, most are used as livestock feed, and the remainder form the basis of purified food ingredients. Although no adverse effects of GM foods in livestock have been detected, the processing of many food ingredients removes most, if not all, plant proteins and DNA. Hence, ingestion of GM food-derived ingredients may not be a sufficient test for health and safety. Second, GM food critics argue that there have been few human clinical trials to directly examine the health effects of most GM foods. One notable exception is Golden Rice 2, which has undergone two small clinical trials. They also say that the toxicity studies that have been completed are performed in animals—primarily rats and mice—and most of these are short-term toxicity studies.

Supporters of GM foods answer these criticisms with several other arguments. The first argument is that short-term toxicity studies in animals are well-established methods for detecting toxins and allergens. The regulatory processes required prior to approval of any GM food demand data from animal toxicity studies. If any negative effects are detected, approval is not given. Supporters also note that several dozen long-term toxicity studies have been published that deal with GM crops such as glyphosate-resistant soybeans and Bt corn, and none of these has shown long-term negative effects on test animals. Those few studies reporting negative effects have been shown to have serious design flaws and their conclusions are considered unreliable. GM food advocates note that human clinical trials are not required for any other food derived from other genetic modification methods such as selective breeding. During standard breeding of plants and animals, genomes may be mutagenized with radiation or chemicals to enhance the possibilities of obtaining a desired phenotype. This type of manipulation has the potential to introduce mutations into genes other than the ones that are directly selected. Also, plants and animals naturally exchange and shuffle DNA in ways that cannot be anticipated. These include interspecies DNA transfers, transposon integrations, and chromosome modifications. These events may result in unintended changes to the physiology of organisms—changes that could potentially be as great as those arising in GM foods.

Environmental Effects

Critics of GM foods point out that GMOs that are released into the environment have both documented and potential consequences for the environment—and hence may indirectly affect human health and safety. GM food advocates argue that these potential environmental consequences can be identified and managed. Here, we will describe two different aspects of GM foods as they may affect the natural environment and agriculture.

1. Emerging herbicide and insecticide resistance. Many published studies report that the planting of herbicide-tolerant and insect-resistant GM crops has reduced the quantities of herbicides and insecticides that are broadly applied to agricultural crops. As a result, the effects of GM crops on the environment have been assumed to be positive. However, these positive effects may be transient, as herbicide and insecticide resistance is beginning to emerge.

   Since glyphosate-tolerant crops were introduced in the mid-1990s, more than 24 glyphosate-resistant weed species have appeared in the United States. Resistant weeds have been found in 18 other countries, and in some cases, the presence of these weeds is affecting crop yields. One reason for the rapid rise of resistant weeds is that farmers have abandoned other weed-management practices in favor of using a single broad-spectrum herbicide. This strong selection pressure has brought the rapid evolution of weed species bearing gene variants that confer herbicide resistance. In response, biotechnology companies are developing new GM crops with tolerance to multiple herbicides. However, scientists argue that weeds will also develop resistance to the use of multiple herbicides, unless farmers vary their weed management practices and incorporate tillage, rotation, and other herbicides along with using the GM crop. Scientists point out that herbicide resistance is not limited to the use of GM crops. Weed populations will evolve resistance to any herbicide used to control them, and the speed of evolution will be affected by the extent to which the herbicide is used.

   Since 1996, more than eight different species of insect pests have evolved some level of resistance to Bt insecticidal proteins. For example, in 2011 scientists reported the first cases of resistance of the western corn rootworm to Bt maize expressing the *cry3Bb1* gene, in maize fields in Iowa. In 2010, scientists from Monsanto detected large numbers of pink bollworms with resistance to the toxin expressed from the *cry1Ac* gene in one variety of Bt cotton. In order to slow down the development of Bt resistance, several strategies are being followed. The first is to develop varieties of GM crops that express two Bt toxins simultaneously. Several of these varieties are already on the market and are replacing varieties that express only one Bt *cry* gene. The second strategy involves the use of “refuges” surrounding fields that grow Bt crops. These refuges contain non-GM crops. Insect pests grow easily within the refuges, which place no evolutionary pressure on the insects for resistance to Bt toxins. The idea is for these non-selected insects to mate with any resistant insects that appear in the Bt crop region of the field. The resulting hybrid offspring will be heterozygous for any resistance gene variant. As long as the resistance gene variant is recessive, the hybrids will be killed by eating the Bt crop. In fields that use refuges and plant GM crops containing two Bt genes, resistance to Bt toxins has been delayed or is absent. As with emerging herbicide resistance, farmers are also encouraged to combine the use of Bt crops with conventional pest control methods.

2. The spread of GM crops into non-GM crops. There have been several documented cases of GM crop plants appearing in uncultivated areas in the United States, Canada, Australia, Japan, and Europe. For example, GM sugar beet plants have been found growing in commercial top soils. GM canola plants have been found growing in ditches and along roadways, railway tracks, and in fill soils, far from the fields in which they were

grown. A 2011 study[1] found “feral” GM canola plants growing in 288 of 634 sample sites along roadways in North Dakota. Of these plants, 41 percent contained the CP4 EPSPS protein (conferring glyphosate resistance), and 39 percent contained the PAT protein (conferring resistance to the herbicide glufosinate). In addition, two of the plants (0.7 percent of the sample) expressed both proteins (resistant to both herbicides). GM plants that express both proteins have not been created by genetic modification and were assumed to have arisen by cross-fertilization of the other two GM crops. The researchers who conducted this survey were not surprised to find GM canola along transportation routes, as seeds are often spilled during shipping. More surprising was the extent of the distribution and the presence of hybridized GM canola plants.

One of the major concerns about the escape of GM crop plants from cultivation is the possibility of **outcrossing** or **gene flow**—the transfer of transgenes from GM crops into sexually compatible non-GM crops or wild plants, conferring undesired phenotypes to the other plants. Gene flow between GM crops and adjacent non-GM crops is of particular concern for farmers who want to market their crops as “GM-free” or “organic” and for farmers who grow seed for planting.

Gene flow of GM transgenes has been documented in GM and non-GM canola as well as sugar beets, and in experiments using rice, wheat, and maize. GM critics often refer to controversial studies about GM outcrossing in Oaxaca, Mexico. In the first study in 2001, it was reported that the local maize crops contained transgenes from Monsanto’s Roundup-Ready and Bt insect-resistant maize. As GM crops were not approved for use in Mexico, it was thought that the transgenes came from maize that had been imported from the United States as a foodstuff, and then had been planted by farmers who were not aware that the seeds were transgenic. Over the next ten years, subsequent studies reported mixed results. In some studies, the transgenes were not detected, and in others, the same transgenes were detected. There is still no consensus about whether gene flow has occurred between the GM and non-GM maize in Mexico.

It is thought that the presence of glyphosate-resistant transgenes in wild plant populations is not likely to be an environmental risk and would confer no positive fitness benefits to the hybrids. The presence of glyphosate-resistant genes in wild populations would, however, make it more difficult to eradicate the plants. This is illustrated in a case of escaped GM bentgrass in Oregon, where it has been difficult to get rid of the plants because it is no longer possible to use the relatively safe herbicide glyphosate. The potential for environmental damage may be greater if the GM transgenes did confer an advantage—such as insect resistance or tolerance to drought or flooding.

In an attempt to limit the spread of transgenes from GM crops to non-GM crops, regulators are considering a requirement to separate the crops so that pollen would be less likely to travel between them. Each crop plant would require different isolation distances to take into account the dynamics of pollen spreading. Several other methods are being considered. For example, one proposal is to make all GM plants sterile using RNAi technology. Another is to introduce the transgenes into chloroplasts. As chloroplasts are inherited maternally, their genomes would not be transferred via pollen. All of these containment methods are in development stages and may take years to reach the market.

ST 4.4 The Future of GM Foods

Over the last 20 years, GM foods have revealed both promise and problems. GM advocates are confident that the next generation of GM food, especially those created using gene-editing technologies, will show even more promising prospects—and may also address many of the problems.

Research is continuing on ways to fortify staple crops with nutrients to address diet problems in poor countries. For example, Australian scientists are adding genes to bananas that will not only provide resistance to Panama disease—a serious fungal disease that can destroy crops—but also increase the levels of beta-carotene and other nutrients, including iron. Other GM crops in the pipeline include plants engineered to resist drought, high salinity, nitrogen starvation, and low temperatures.

Researchers are also devising more creative ways to protect plants from insects and diseases. One intriguing project involves introducing into wheat a gene that encodes a pheromone that acts as a chemical alarm signal to aphids. If successful, this approach could protect the wheat plants from aphids without using toxins. Another project involves cassava, which is a staple crop for many Africans and is afflicted by two viral diseases—cassava mosaic virus and brown streak virus—that stunt growth and cause root rot. Although some varieties of cassava are resistant to these viruses, the life cycle of cassava is so long that it would be difficult to introduce resistance into other varieties using conventional breeding techniques. Scientists plan to transform plants with genes from resistant cassava. This type of cisgenic gene transfer is more comparable to traditional breeding than transgenic techniques.

[1] Schafer, M. G. et al. (2011). *PLoS One* 6:e25736.

In the future, GM foods will likely include additional GM animals. In one project, scientists have introduced a DNA sequence into chickens that protects the birds from spreading avian influenza. The sequence encodes a hairpin RNA molecule with similarity to a normal viral RNA that binds to the viral polymerase. The presence of the hairpin RNA inhibits the activity of the viral polymerase and interferes with viral propagation. If this strategy proves useful *in vivo*, the use of these GM chickens would not only reduce the incidence of avian influenza in poultry production, but also reduce the transmissibility of avian influenza viruses to humans.

Although these and other GM foods show promise for increasing agricultural productivity and decreasing disease, the political pressure from anti-GM critics remains a powerful force. An understanding of the science behind these technologies will help us all to evaluate the future of GM foods.

Selected Readings and Resources

Journal Articles

Butler, D. (2012). Hyped GM maize study faces growing scrutiny. *Nature* 490:158.

Cressey, D. (2013). A new breed. *Nature* 497:27–29.

Enserink, M. (2008). Tough lessons from Golden Rice. *Science* 320:468–471.

Gassmann, A. J., et al. (2011). Field-evolved resistance to Bt maize by western corn rootworm. *PLoS One* 6(7):e22629.

Gilbert, N. (2013). A hard look at GM crops. *Nature* 497:21–26.

Ledford, H. (2015). Transgenic salmon leaps to the dinner table. *Nature* 527:417–418.

Oke, K. B., et al. (2013). Hybridization between genetically modified Atlantic salmon and wild brown trout reveals novel ecological interactions. *Proc. R. Soc. B.* 280:20131047.

Paine, J. A., et al. (2005). Improving the nutritional value of Golden Rice through increased pro-vitamin A content. *Nature Biotech.* 23(4):482–487.

Schafer, M. G., et al. (2011). The establishment of genetically engineered canola populations in the US. *PLoS One* 6(10):e25736.

Waltz, E. (2016). Gene-edited CRISPR mushroom escapes US regulation. *Nature* 532:293.

Whitty, C. J. M., et al. (2013). Africa and Asia need a rational debate on GM crops. *Nature* 497:31–33.

Web Sites

The International Rice Research Institute. Golden rice. http://irri.org/golden-rice

International Service for the Acquisition of Agri-Biotech Applications. GM approval databases and information. http://isaaa.org

U.S. Department of Agriculture, Economic Research Service. Adoption of genetically engineered crops in the US. http://www.ers.usda.gov/data-products/adoption-of-genetically-engineered-crops-in-the-us.aspx#.UdmLqW3fJyI

World Health Organization. Twenty questions on genetically modified foods. http://www.who.int/foodsafety/areas_work/food-technology/faq-genetically-modified-food/en/

Review Questions

1. How do genetically modified organisms compare with organisms created through selective breeding?
2. Can current GM crops be considered as transgenic or cisgenic? Why?
3. Of the approximately 200 GM crop varieties that have been developed, only a few are widely used. What are these varieties, and how prevalent are they?
4. How does glyphosate work, and how has it been used with GM crops to increase agricultural yields?
5. Describe the mechanisms by which the Cry proteins from *Bacillus thuringiensis* act as insecticides.
6. What measures have been taken to alleviate vitamin A deficiencies in developing countries? To date, how successful have these strategies been?
7. What is Golden Rice 2, and how was it created?
8. Describe how plants can be transformed using biolistic methods. How does this method compare with *Agrobacterium tumefaciens*–mediated transformation?
9. How do positive and negative selection techniques contribute to the development of GM crops?
10. Describe how the Roundup-Ready soybean variety was developed, and what genes were used to transform the soybean plants.

Discussion Questions

1. What are the laws regulating the development, approval, and use of GM foods in your region and nationally?
2. Do you think that foods containing GM ingredients should be labeled as such? What would be the advantages and disadvantages to such a strategy?
3. One of the major objections to GM foods is that they may be harmful to human health. Do you agree or disagree, and why?

SPECIAL TOPICS IN MODERN GENETICS 5

Gene Therapy

Although drug treatments can be effective in controlling symptoms of genetic disorders, the ideal outcome of medical treatment is to cure a disease. This is the goal of **gene therapy**—the delivery of therapeutic genes into a patient's cells to correct genetic disease conditions caused by a faulty gene or genes. The earliest attempts at gene therapy focused on the delivery of normal, *therapeutic copies* of a gene to be expressed in such a way as to override or negate the effects of the disease gene and thus minimize or eliminate symptoms of the genetic disease. But in recent years newer methods for inhibiting or silencing defective genes have shown promise. However, no approach to gene therapy has generated more excitement for its potential than gene-editing applications involving CRISPR-Cas. We will consider several recent examples of gene editing for targeted removal of defective genes.

Gene therapy is one of the goals of **translational medicine**—taking a scientific discovery, such as the identification of a disease-causing gene, and translating the finding into an effective therapy, thus moving from the laboratory bench to a patient's bedside to treat a disease. In theory, the delivery of a therapeutic gene is rather simple, but in practice, gene therapy has been very difficult to execute. In spite of over 25 years of trials, this field has not lived up to its expectations. However, gene therapy is currently experiencing a fast-paced resurgence, with several high-profile new successes and potentially exciting new technologies sitting on the horizon. It is hoped that gene therapy will soon become part of mainstream medicine. The treatment of a human genetic disease by gene therapy is the ultimate application of genetic technology. In this Special Topic chapter we will explore how gene therapy is executed, and we will highlight selected examples of successes and failures as well as discuss new approaches to gene therapy. Finally, we will consider ethical issues regarding gene therapy.

> "The treatment of a human genetic disease by gene therapy is the ultimate application of genetic technology."

ST 5.1 What Genetic Conditions Are Candidates for Treatment by Gene Therapy?

Two essential criteria for gene therapy are that the gene or genes involved in causing a particular disease have been identified and that the gene can be cloned or synthesized in a laboratory. As a result of the Human Genome Project, the identification of human disease genes and their specific DNA sequences has greatly increased the number of candidate genes for gene therapy trials. Almost all of the early gene therapy trials and most gene therapy approaches have focused on treating conditions caused by a single gene.

The cells affected by the genetic condition must be readily accessible for treatment by gene therapy. For example, blood disorders such as leukemia, hemophilia, and other conditions have been major targets of gene therapy because it is relatively routine to manipulate blood cells outside of the body and return them to the body in comparison to treating cells in the brain and spinal cord, skeletal or cardiac muscle, and organs with heterogeneous populations of cells such as the pancreas.

In the past decade, every major category of genetic diseases has been targeted by gene therapy (**Figure ST 5.1**). A majority of recently approved clinical trials are for cancer treatment. Gene therapy approaches are also currently being investigated for the treatment of hereditary blindness; hearing loss; neurodegenerative diseases including Alzheimer disease, Parkinson disease, and amyotrophic lateral sclerosis (ALS); cardiovascular disease; muscular dystrophy; hemophilia; and infectious diseases, such as HIV; among many other conditions, including depression and drug and alcohol addiction. Worldwide, over 2300 approved gene therapy clinical trials have occurred or recently been initiated.

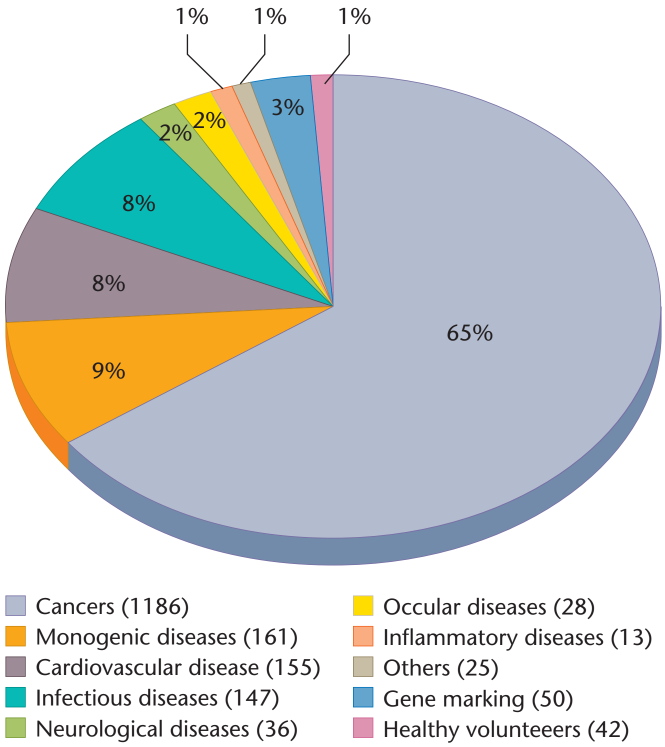


FIGURE ST 5.1 Graphic representation of different genetic conditions being treated by gene therapy clinical trials worldwide. Notice that cancers are the major target for treatment.

In the United States, proposed gene therapy clinical trials must first be approved by review boards at the institution where they will be carried out, and then the protocols must be approved by the Food and Drug Administration (FDA).

ST 5.2 How Are Therapeutic Genes Delivered?

In general, there are two broad approaches for delivering therapeutic genes to a patient being treated by gene therapy, *ex vivo gene therapy* and *in vivo gene therapy* (**Figure ST 5.2**). In *ex vivo* gene therapy, cells from a person with a particular genetic condition are removed, treated in a laboratory by adding either normal copies of a therapeutic gene or a DNA or RNA sequence that will inhibit expression of a defective gene, and then these cells are transplanted back into the person. Genetically altered cells treated in this manner can be transplanted back into the patient without fear of immune system rejection because these cells were derived from the patient initially.

In vivo gene therapy does not involve removal of a person's cells. Instead, therapeutic DNA is introduced directly into affected cells of the body. One of the major challenges of *in vivo* gene therapy is restricting the delivery of therapeutic genes to only the intended tissues and not to all tissues throughout the body.

Viral Vectors for Gene Therapy

For both *in vitro* and *ex vivo* approaches, the key to successful gene therapy is having a delivery system to transfer genes into a patient's cells. Because of the relatively large molecular size and electrically charged properties of DNA, most human cells do not take up DNA easily. Therefore, delivering therapeutic DNA molecules into human cells is challenging. Since the early days of gene therapy, genetically engineered viruses as vectors have been the main tools for delivering therapeutic genes into human cells. Viral vectors for gene therapy are engineered to carry therapeutic DNA as their payload so that the virus infects target cells and delivers the therapeutic DNA without causing damage to cells.

In a majority of gene therapy trials around the world, scientists have used genetically modified *retroviruses* as vectors. Recall from Chapter 10 that retroviruses such as HIV contain an RNA genome that scientists use as a template for the synthesis of a complementary DNA molecule. **Retroviral vectors** are created by removing replication and disease-causing genes from the virus and replacing them with a cloned human gene. After the altered RNA has been packaged into the virus, the recombinant viral vector containing the therapeutic human gene is used to infect a patient's cells. Technically, virus particles are carrying RNA copies of the therapeutic gene. Once inside a cell, the virus cannot replicate itself, but the therapeutic RNA is reverse transcribed into DNA, which enters the nucleus of cells and *integrates* into the genome of the host cells' chromosome. If the inserted therapeutic gene is properly expressed, it produces a normal gene product that may be able to ameliorate the effects of the mutation carried by the affected individual.

One advantage of retroviral vectors is that they provide long-term expression of delivered genes because they integrate the therapeutic gene into the genome of the patient's cells. But a major problem with retroviral vectors is that they have produced random, unintended alterations in the genome, in some cases due to *insertional mutations*. Retroviral vectors generally integrate their genome into the host-cell genome at random sites. Thus, there is the potential for retroviral integration that randomly inactivates genes in the genome or gene-regulatory regions such as a promoter sequence.

In many early gene therapy trials, **adenovirus vectors** were the retrovirus vector of choice. An advantage of these vectors is that they are capable of carrying large therapeutic genes. But because many humans produce antibodies to adenovirus vectors they can mount

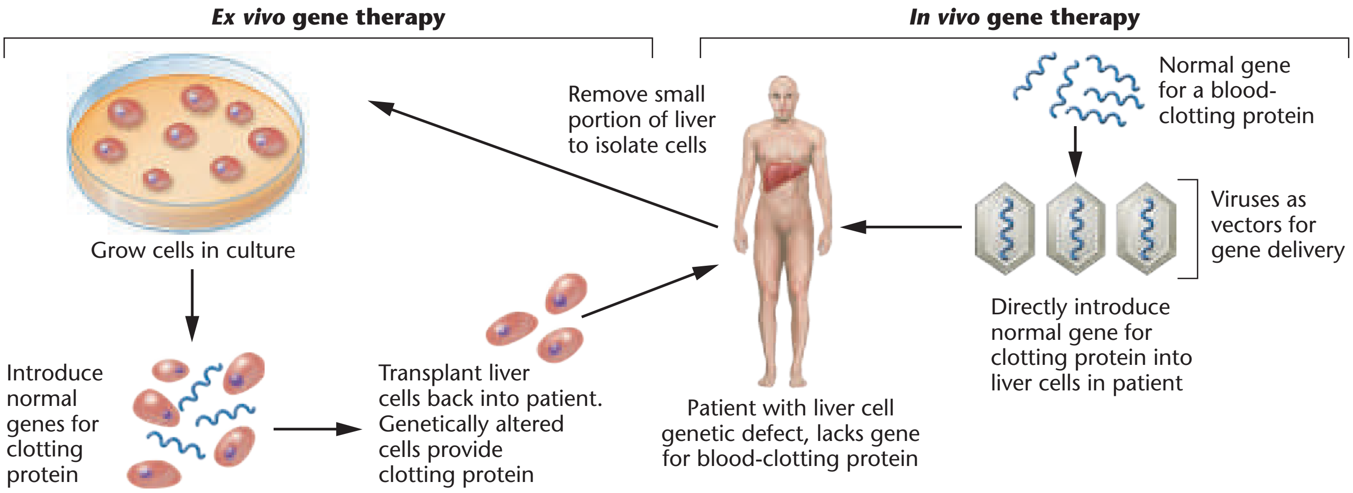

Ex vivo gene therapy
In vivo gene therapy
Remove small portion of liver to isolate cells
Normal gene for a blood-clotting protein
Viruses as vectors for gene delivery
Grow cells in culture
Directly introduce normal gene for clotting protein into liver cells in patient
Introduce normal genes for clotting protein
Transplant liver cells back into patient. Genetically altered cells provide clotting protein
Patient with liver cell genetic defect, lacks gene for blood-clotting protein

FIGURE ST 5.2 *Ex vivo* and *in vivo* gene therapy for a patient with a liver disorder. *Ex vivo* gene therapy involves isolating cells from the patient, introducing normal copies of a therapeutic gene (encoding a blood-clotting protein in this example) into these cells, and then returning cells to the body where they will produce the required clotting protein. *In vivo* approaches involve introducing DNA directly into cells while they are in the body.

immune reactions that render the virus and its therapeutic gene ineffective or cause significant side effects to the patient. A related virus called **adeno-associated virus (AAV)** is now widely used as a gene therapy vector [**Figure ST 5.3(a)**]. In its native form, AAV infects about 80–90 percent of humans during childhood, causing symptoms associated with the common cold. Disabled forms of AAV are popular for gene therapy because the virus is nonpathogenic, so it usually does not elicit a major response from the immune system of treated patients. AAV also does not typically integrate into the host-cell genome, so there is little risk of the insertional mutations that have plagued retroviruses, although modified forms of AAV have been used to deliver genes to specific sites on individual chromosomes. Most forms of AAV deliver genes into the host-cell nucleus where it forms small hoops of DNA called *episomes* that are expressed under the control of promoter sequences contained within the viral genome. But because therapeutic DNA delivered by AAV does not usually become incorporated into the genome, it is not replicated when host cells divide. This is fine for certain cells in the brain or the retina that do not divide, but treating rapidly dividing cells typically requires repeated, ongoing applications to be successful [Figure ST 5.3(a)].

Work with **lentivirus vectors** is an active area of gene therapy research [**Figure ST 5.3(b)**]. Lentivirus is a retrovirus that can accept relatively large pieces of genetic material. Another positive feature of lentivirus is that it is capable of infecting nondividing cells, whereas other viral vectors often infect cells only when they are dividing. It is still not possible to control where lentivirus integration occurs in the host-cell genome, but the virus does not appear to gravitate toward gene-regulatory regions the way that other retroviruses do. Thus the likelihood of causing insertional mutations appears to be much lower than for other vectors.

The human immunodeficiency virus (HIV) responsible for acquired immunodeficiency syndrome (AIDS) is a type of lentivirus. It may surprise you that HIV could

BOX 1
ClinicalTrials.gov

One of the best resources on the Web for learning about ongoing clinical trials, including current gene therapy trials, is ClinicalTrials.gov. The site can easily be searched to find a wealth of resources about ongoing gene therapy trials throughout the United States that are of interest to you. To find a gene therapy clinical trial, use the "Search for Studies" box and type in the name of a disease and "gene therapy." This search string will take you to a page listing active gene therapy clinical trials, with links to detailed information about the trial.

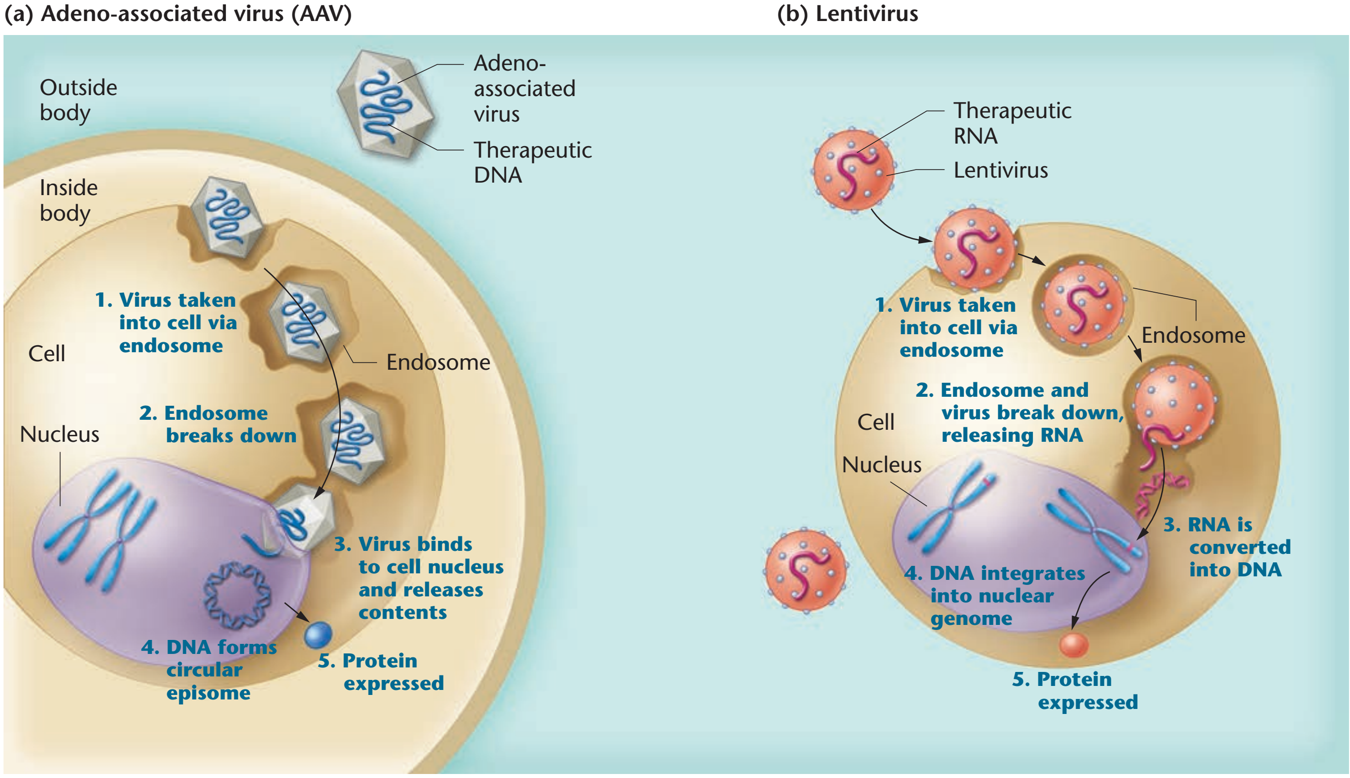

(a) Adeno-associated virus (AAV)
Outside body
Adeno-associated virus
Therapeutic DNA
Inside body
1. Virus taken into cell via endosome
Cell
Endosome
Nucleus
2. Endosome breaks down
3. Virus binds to cell nucleus and releases contents
4. DNA forms circular episome
5. Protein expressed
(b) Lentivirus
Therapeutic RNA
Lentivirus
1. Virus taken into cell via endosome
Endosome
2. Endosome and virus break down, releasing RNA
Cell
Nucleus
3. RNA is converted into DNA
4. DNA integrates into nuclear genome
5. Protein expressed

FIGURE ST 5.3 Delivering therapeutic genes. (a) Nonintegrating viruses such as modified adeno-associated virus (AAV) deliver therapeutic genes without integrating them into the genome of target cells. Delivered DNA resides as minichromosomes (episomes), but over time as cells divide, these nonintegrating hoops of DNA are gradually lost. (b) Integrating viruses include lentivirus, an RNA retrovirus that delivers therapeutic genes into the cytoplasm where reverse transcriptase converts RNA into DNA. DNA then integrates into the genome, ensuring that therapeutic DNA will be passed into daughter cells during cell division.

be used as a vector for gene therapy. For any viral vector, scientists must be sure that the vector has been genetically engineered to render it inactive. Modified forms of HIV, strains lacking the genes necessary for reconstitution of fully functional viral particles, are being used for gene therapy trials. HIV has evolved to infect certain types of T lymphocytes (T cells) and macrophages, making it a good vector for delivering therapeutic genes into the bloodstream.

Nonviral Delivery Methods

Scientists continue to experiment with various *in vivo* and *ex vivo* strategies for trying to deliver so-called naked DNA into cells without the use of viral vectors. Nonviral methods include chemically assisted transfer of genes across cell membranes, nanoparticle delivery of therapeutic genes, and fusion of cells with artificial lipid vesicles called *liposomes*. Short-term expression of genes through "gene pills" is being explored, whereby a pill delivers therapeutic DNA to the intestines where the DNA is absorbed by cells that express the therapeutic protein and secrete it into the bloodstream.

Stem Cells for Delivering Therapeutic Genes

Increasingly, viral and nonviral vectors are being used to deliver therapeutic genes into **stem cells,** usually *in vitro*, and then the stem cells are either introduced into the patient or differentiated *in vitro* into mature cell types before being transplanted into a patient. In particular, *hematopoietic stem cells (HSCs)* which are found in bone marrow and give rise to blood cells, are widely used for gene therapy in adults and children. There are many advantages to using HSCs: they are easily accessible and often taken from the marrow of the patient to be treated to avoid complications with tissue rejection by the immune system when they are reintroduced; they replicate quickly *in vitro*; they are fairly long lived; and they differentiate into both red blood cells and white blood cells (leukocytes). And as you will also learn, stem cells are being used in CRISPR-Cas and gene-silencing approaches, thus demonstrating their value for different gene therapy applications.

As this edition of Concepts of Genetics (it) went to press, scientists reported remarkable results using gene therapy to deliver the LAMB3 [it] gene into stem cells to restore the epidermis to a boy who lost more than 80% of his skin due to a skin blistering disease called epidermolysis bullosa. Refer to the reference by Hirsch et al. at the end of this chapter to learn more about this late-breaking and quite extraordinary success.

ST 5.3 The First Successful Gene Therapy Trial

In 1990 the FDA approved the first human gene therapy trial, which began with the treatment of a young girl named Ashanti DeSilva, who has a heritable disorder called **adenosine deaminase severe combined immunodeficiency (ADA-SCID).** ADA-SCID is a condition affecting approximately 1-9 out of every 1 million live births. Individuals with SCID have no functional immune system and usually die from what would normally be minor infections. Ashanti has an autosomal form of SCID caused by a mutation in the gene encoding the enzyme *adenosine deaminase*. Her gene therapy began when clinicians isolated some of her white blood cells, called T cells [**Figure ST 5.4**]. These cells, which are key components of the immune system, were mixed with a retroviral vector carrying an inserted copy of the normal *ADA* gene. The virus infected many of the T cells, and a normal copy of the *ADA* gene was inserted into the genome of some T cells.

After being mixed with the vector, the T cells were grown in the laboratory and analyzed to make sure that the transferred *ADA* gene was expressed (Figure ST 5.4). Then a billion or so genetically altered T cells were injected into Ashanti's bloodstream. Repeated treatments were required to produce a sufficient number of functioning T cells. In addition, Ashanti also periodically received injections of purified ADA protein throughout this process, so the exact effects of gene therapy were difficult to discern. Ashanti continues to receive supplements of the ADA enzyme to allow her to lead a normal life.

Subsequent gene therapy treatments for SCID have focused on using bone marrow stem cells called hematopoietic stem cells (HSCs), and *in vitro* approaches to repopulate the number of ADA-producing T cells. In the past decade alone, over 100 people (mostly children) have received gene therapy for ADA-SCID, and most have been

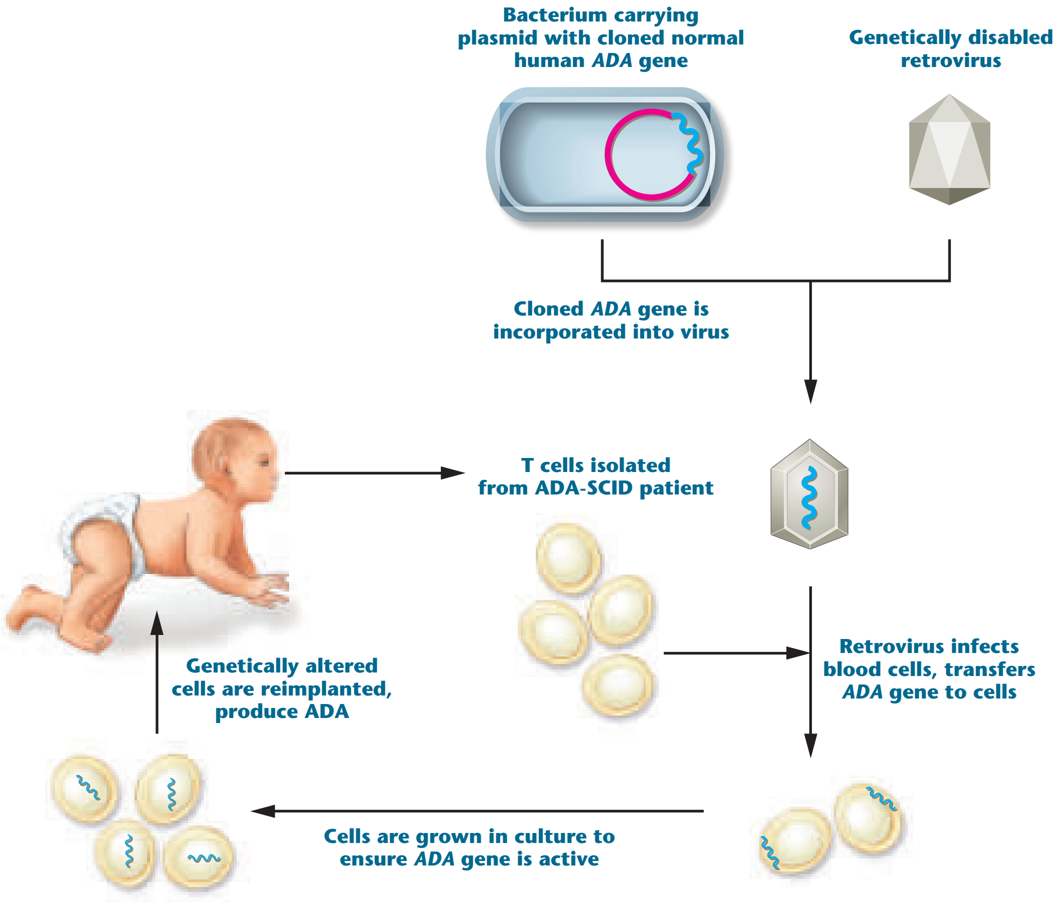

Bacterium carrying plasmid with cloned normal human *ADA* gene
Genetically disabled retrovirus
Cloned *ADA* gene is incorporated into virus
T cells isolated from ADA-SCID patient
Retrovirus infects blood cells, transfers *ADA* gene to cells
Cells are grown in culture to ensure *ADA* gene is active
Genetically altered cells are reimplanted, produce ADA

FIGURE ST 5.4 The first successful gene therapy trial. To treat ADA-SCID using gene therapy, a cloned human *ADA* gene was transferred into a viral vector, which was then used to infect white blood cells removed from the patient. The transferred *ADA* gene was incorporated into a chromosome, after which the cells were cultured to increase their numbers. Finally, the cells were inserted back into the patient, where they produce ADA, allowing the development of an immune response.

treated successfully and are disease-free. ADA-SCID treatment is still considered the most successful example of gene therapy.

ST 5.4 Gene Therapy Setbacks

From 1990 to 1999, more than 4000 people underwent gene therapy for a variety of genetic disorders. These trials often failed and thus led to a loss of confidence in gene therapy. In the United States, optimism for gene therapy plummeted even further in 1999 when teenager Jesse Gelsinger died while undergoing a test for the safety of gene therapy to treat a liver disease called *ornithine transcarbamylase* (OTC) deficiency. Large numbers of adenovirus vectors bearing the *OTC* gene were injected into his hepatic artery. The vectors were expected to target his liver, enter liver cells, and trigger the production of OTC protein. In turn, it was hoped that the OTC protein might correct his genetic defect and cure him of his liver disease.

Researchers had previously treated 17 people with the therapeutic virus, and early results from these patients were promising. But as the 18th patient, Jesse Gelsinger, within hours of his first treatment, developed a massive immune reaction. He developed a high fever, his lungs filled with fluid, multiple organs shut down, and he died four days later of acute respiratory failure. Jesse's severe response to the adenovirus may have resulted from how his body reacted to a previous exposure to the virus used as the vector for this protocol.

In the aftermath of the tragedy, several government and scientific inquiries were conducted. Investigators learned that in the clinical trial scientists had not reported other adverse reactions to gene therapy and that some of the scientists were affiliated with private companies that could benefit financially from the trials. It was determined that serious side effects seen in animal studies were not explained to patients during informed-consent discussions. The FDA subsequently scrutinized gene therapy trials across the country, halted a number of them, and shut down several gene therapy programs. Other groups voluntarily suspended their gene therapy studies. Tighter restrictions on clinical trial protocols were imposed to correct some of the procedural problems that emerged from the Gelsinger case. Jesse's death had dealt a severe blow to the struggling field of gene therapy—a blow from which it was still reeling when a second tragedy hit.

The outlook for gene therapy brightened momentarily in 2000, when a group of French researchers reported what was hailed as the first large-scale success in gene therapy. Children with a fatal X-linked form of ADA-SCID (X-SCID, also known as "bubble boy" disease) developed functional immune systems after being treated with a retroviral vector carrying a normal gene. But elation over this study soon turned to despair, when it became clear that 5 of the 20 patients in two different trials developed leukemia as a direct result of their therapy. One of these patients died as a result of the treatment, while the other four went into remission from the leukemia. In two of the children examined, their cancer cells contained the retroviral vector, inserted near a gene called *LMO2*. This *insertional mutation* activated the *LMO2* gene, causing uncontrolled white blood cell proliferation and development of leukemia. The FDA immediately halted 27 similar gene therapy clinical trials, and once again gene therapy underwent a profound reassessment.

On a positive note, long-term survival data from trials in the UK to treat X-SCID and ADA-SCID using HSCs from the patients' bone marrow for gene therapy have shown that 14 of 16 children have had their immune system restored at least 9 years after the treatment. These children formerly had life expectancies of less than 20 years. In the past 5 years alone, more than 40 patients have been treated for ADA-SCID in three well-developed programs in Italy, the United Kingdom, and the United States. All individuals treated have survived, and 75 percent of those treated are disease-free.

Problems with Gene Therapy Vectors

Critics of gene therapy have berated researchers for undue haste, conflicts of interest, sloppy clinical trial management, and for promising much but delivering little. Most of the problems associated with gene therapy, including the Jesse Gelsinger case and the French X-SCID trial, have been traced to the viral vectors used to transfer therapeutic genes into cells. These vectors have been shown to have several serious drawbacks.

- First, integration of retroviral genomes, including the human therapeutic gene into the host cell's genome, occurs only if the host cells are replicating their DNA. In the body, only a small number of cells in any tissue are dividing and replicating their DNA.
- Second, the injection of massive quantities of most viral vectors, but particularly adenovirus vectors, is capable of causing an adverse immune response in the patient, as happened in Jesse Gelsinger's case.
- Third, insertion of viral genomes into host chromosomes can activate or mutate an essential gene, as in the case of the French patients. Viral integrase, the enzyme that allows for viral genome integration into the host genome, interacts with chromatin-associated proteins, often steering integration toward transcriptionally active genes.

- Fourth, AAV vectors cannot carry DNA sequences larger than about 5 kb, and retroviruses cannot carry DNA sequences much larger than 10 kb. Many human genes exceed the 5—10 kb size range.
- Finally, there is a possibility that a fully infectious virus could be created if the inactivated vector were to recombine with another unaltered viral genome already present in the host cell.

To overcome these problems, new viral vectors and strategies for transferring genes into cells are being developed in an attempt to improve the action and safety of vectors. No new technology has had a greater impact on gene therapy than gene targeting, especially by CRISPR-Cas. Fortunately, gene therapy has experienced a resurgence in part because of several promising new trials and successful treatments.

ST 5.5 Recent Successful Trials by Conventional Gene Therapy Approaches

Treating Retinal Blindness

In recent years, patients being treated for blindness have greatly benefited from gene therapy approaches. Congenital retinal blinding conditions affect about 1 in 2000 people worldwide, many of which are the result of a wide range of genetic defects. Over 165 different genes have been implicated in various forms of retinal blindness. Currently there are over two dozen active gene therapy trials for at least 10 different retinal diseases.

Successful gene therapy has been achieved in subsets of patients with **Leber congenital amaurosis (LCA),** a degenerative disease of the retina that affects 1 in 50,000 to 1 in 100,000 infants each year and causes severe blindness. Gene therapy treatments for LCA were originally pioneered in dogs. Based on the success of these treatments, the protocols were adapted and applied to human gene therapy trials.

LCA is caused by alterations to photoreceptor cells (rods and cones), light-sensitive cells in the retina, due to 18 or more genes. One gene in particular, *RPE65*, has been the gene therapy target of choice. The protein product of the *RPE65* gene metabolizes retinol, which is a form of vitamin A that allows the rod and cone cells of the retina to detect light and transmit electrical signals to the brain. In one of the earliest trials, young adult patients with defects in the *RPE65* gene were given injections of the normal gene incorporated into an AAV vector. Several months after a single treatment, many adult patients, while still legally blind, could detect light, and some of them could read lines of an eye chart. This treatment approach for LCA was based on injecting AAV-carrying *RPE65* at the back of the eye directly under the retina (**Figure ST 5.5**). The therapeutic gene enters about 15 to 20 percent of cells in the retinal pigment epithelium, the layer of cells just beneath the visual cells of the retina. Adults treated by this approach have shown substantial improvements in a variety of visual functions tests, but the greatest improvement has been demonstrated in children, all of whom have gained sufficient vision to allow them to be ambulatory. Researchers think the success in children has occurred because younger patients have not lost as many photoreceptor cells as older patients. The FDA is expected to approve this gene therapy approach in early 2018.

Over two dozen gene therapy trials have been completed or are ongoing for various forms of blindness, including age-related degenerative causes of blindness. Because of the small size of the eye and the relatively small number of cells that need to be treated, the prospects for gene therapy to become routine treatment for eye disorders appears to be very good. Retinal cells are also very long lived; thus, AAV delivery approaches can be successful for long periods of time even if the gene does not integrate.

Successful Treatment of Hemophilia B

A very encouraging gene therapy trial in England successfully treated a small group of adults with hemophilia B, a blood disorder caused by a deficiency in the coagulation protein human factor IX. This, and other similar trials, are based largely on approaches derived from successful gene

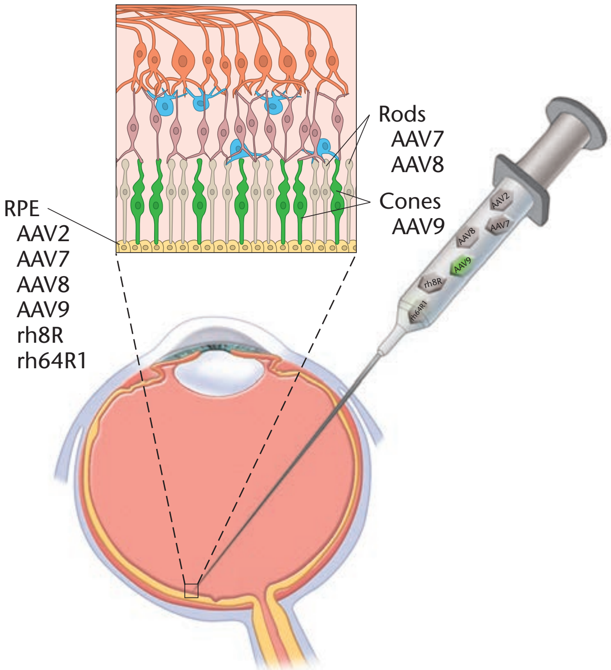


FIGURE ST 5.5 Treatment of retinal blindness. Illustration of AAV delivery of specific genes targeting individual cell types of the retina, rods, cones, and the layer of retinal pigment epithelial (RPE) cells. The basic approach shown here was the delivery method used to successfully treat LCA.

therapy to treat hemophilia B in dogs. Currently, most hemophilia B patients are treated several times each week with infusions of concentrated doses of the factor IX protein. In the gene therapy trial, six adult patients received, *in vivo*, a single dose of an adenovirus vector (AAV8) carrying normal copies of the human factor IX gene introduced into liver cells. Of six patients treated, four were able to stop factor IX infusion treatments after the gene therapy trial. Several other trials of this AAV treatment approach are under way, and expectations are high that a gene therapy cure for hemophilia B is close to becoming a routine reality.

HIV as a Vector Shows Promise in Recent Trials

Researchers at the University of Paris and Harvard Medical School reported that two years after gene therapy treatment for β-**thalassemia**, a blood disorder involving the β-globin gene that reduces the production of hemoglobin, a young man no longer needed transfusions and appeared to be healthy. A modified, disabled HIV was used to carry a copy of the normal β-globin gene. Although this trial resulted in activation of the growth factor gene called *HMGA2*, reminiscent of what occurred in the French X-SCID trials, activation of the transcription factor did not result in an overproduction of hematopoietic cells or create a condition of preleukemia.

In 2013, researchers at the San Raffaele Telethon Institute for Gene Therapy in Milan, Italy, first reported two studies using lentivirus vectors derived from HIV in combination with HSCs to successfully treat children with either **metachromatic leukodystrophy (MLD)** or **Wiskott–Aldrich syndrome (WAS).** MLD is a neurodegenerative disorder affecting storage of enzymes in lysosomes and is caused by mutation in the arylsulfatase A (*ARSA*) gene that results in an accumulation of fats called sulfatides. These are toxic to neurons, causing progressive loss of the myelin sheath (demyelination) surrounding neurons in the brain, leading to a loss of cognitive functions and motor skills. There is no cure for MLD. Children with MLD appear healthy at birth but eventually develop MLD symptoms.

Researchers used an *ex vivo* approach with a lentivirus vector to introduce a functional *ARSA* gene into bone marrow—derived HSCs from each patient and then infused treated HSCs back into patients. Four years after the start of a trial involving 10 patients with MLD, data from six patients analyzed 18 to 24 months after gene therapy indicated that the trials are safe and effective. Treatment halted disease progression as determined by magnetic resonance images of the brain and through tests of cognitive and motor skills. Patients with MLD in the first group treated have already lived past the expected lifetime normally associated with this disease. Additional patients are now being treated. This approach took over 15 years of research and a team of over 70 people, including researchers and clinicians, which is indicative of the teamwork approach typical of gene therapy trials.

The trial was technically complicated because it required that HSCs travel through the bloodstream and release the ARSA protein that is taken up into neurons. A major challenge was to create enough engineered cells to produce a sufficient quantity of therapeutic ARSA protein to counteract the neurodegenerative process.

Similar results were reported for treating patients with WAS, an X-linked condition resulting in defective platelets that make patients more vulnerable to infections, frequent bleeding, autoimmune diseases, and cancer. Genome sequencing of MLD and WAS patients treated in these trials showed no evidence of genome integration near oncogenes. Similarly, patients showed no evidence of HSC overproduction, suggesting that this lentivirus delivery protocol produced a safe and stable delivery of the therapeutic genes.

BOX 2

Glybera: The First Commercial Gene Therapy to be Approved in the West Lasted Only 5 Years

In late 2012, a gene therapy product called Glybera® (alipogene tiparvovec), developed by Amsterdam-based company uniQure, made history when the European Medicines Agency of the European Union approved it as the first gene therapy trial to win commercial approval in the Western world. Glybera is an AAV vector system for delivering therapeutic copies of the LPL gene to treat patients with a rare disease called lipoprotein lipase deficiency (LPLD, also called familial hyperchylomicronemia). LPLD patients have high levels of triglycerides in their blood. Elevated serum triglycerides are toxic to the pancreas and cause a severe form of pancreatic inflammation called pancreatitis. Because, the U.S. FDA requested additional clinical trials before it would consider approval of Glybera, uniQure discontinued plans to seek approval by the FDA. The success of Glybera trials in Europe signaled what many researchers hoped would be the beginning of many new gene therapy approvals in Europe and the United States. But despite its promise, by 2017, Glybera had failed to be widely used by any European country and since inception had only been used in one patient. Glybera failed, in part, because at a cost of over $1 million per treatment, it was one of the most expensive drugs in history. Thus uniQure announced it would not renew European marketing authorization after October 2017.

ST 5.6 Gene Editing Approaches to Gene Therapy

The gene therapy approaches and examples we have highlighted thus far have focused on the addition of a therapeutic gene that functions along with the defective gene. However, rapid progress is being made with **gene editing**—the removal, correction, and/or replacement of a mutated gene. Gene editing by CRISPR-Cas, in particular, has shown great potential and provided renewed optimism for scientists and physicians involved in gene therapy as well as patients.

DNA-Editing Nucleases

For nearly 20 years, scientists have been working on modifications of restriction enzymes and other nucleases to engineer proteins capable of editing the genome with precision, including the ability to edit one or a few bases or to replace specific genes. The concept is to combine a nuclease with a sequence-specific DNA binding domain that can be precisely targeted for digestion. In 1996 researchers fused DNA-binding proteins with a zinc-finger motif and DNA cutting domain from the restriction enzyme *Fok*I to create enzymes called **zinc-finger nucleases** (**ZFNs**; **Figure ST 5.6**). The zinc-finger motif is found in many transcription factors and consists of a cluster of two cysteine and two histidine residues that bind zinc atoms and interact with specific DNA sequences. By coupling zinc-finger motifs to DNA cutting portions of a polypeptide, ZFNs provide a mechanism for modifying sequences in the genome in a sequence-specific targeted way.

The DNA-binding domain of the ZFN can be engineered to attach to any sequence in the genome. The zinc fingers bind with a spacing of 5–7 nucleotides, and the nuclease domain of the ZFN cleaves between the binding sites.

Another category of DNA-editing nucleases called **transcription activator-like effector nucleases (TALENs)** was created by adding a DNA-binding motif identified in transcription factors from plant pathogenic bacteria known as transcription activator-like effectors (TALEs) to nucleases to create TALENs. TALENs also cleave as dimers. The DNA-binding domain is a tandem array of amino acid repeats, with each TALEN repeat binding to a specific single base pair. The nuclease domain then cuts the sequence between the dimers, a stretch that spans about 13 bp.

ZFNs and TALENs have shown promise in animal models and cultured cells for gene replacement approaches that involve removing a defective gene from the genome. ZFNs, TALENs, and CRISPR-Cas as we will discuss shortly, all create double-stranded breaks in the DNA and then are mended by either nonhomologous end joining or homologous recombination. These enzymes can create site-specific double-stranded cleavage in the genome. When coupled with certain integrases, ZFNs and TALENs may lead to gene editing by cutting out defective sequences and using recombination to introduce homologous sequences into the genome that replace defective sequences. Although this technology has not yet advanced sufficiently for reliable use in humans, there have been several promising trials.

For example, ZFNs are actively being used in clinical trials for treating patients with HIV. Scientists are exploring ways to deliver immune system–stimulating genes that could make individuals resistant to HIV infection or cripple the virus in HIV-positive persons. In 2007, Timothy Brown, a 40-year-old HIV-positive American, had a relapse of acute myeloid leukemia and received a stem cell transplant. Because he was HIV-positive, Brown's physician selected a donor with a mutation in both copies of the *CCR5* gene, which encodes an HIV co-receptor carried on the surface of T cells to which HIV must bind to enter T cells (specifically CD4+ cells). People with naturally occurring mutations in both copies of the *CCR5* gene are resistant to most forms of HIV. Brown relapsed again and received another stem cell transplant from the *CCR5*-mutant donor. Eventually, the cancer was contained, and by 2010, levels of HIV in his body were still undetectable even though he was no longer receiving immune-suppressive treatment. Brown is generally considered to be the first person to have been cured of an HIV infection.

This example encouraged researchers to press forward with a gene therapy approach to modify the *CCR5* gene of HIV patients. In the first gene editing trial to treat people with HIV, T cells were removed from HIV-positive men, and ZFNs were used to disrupt the *CCR5* gene *ex vivo*. The modified cells were then reintroduced into patients. In five

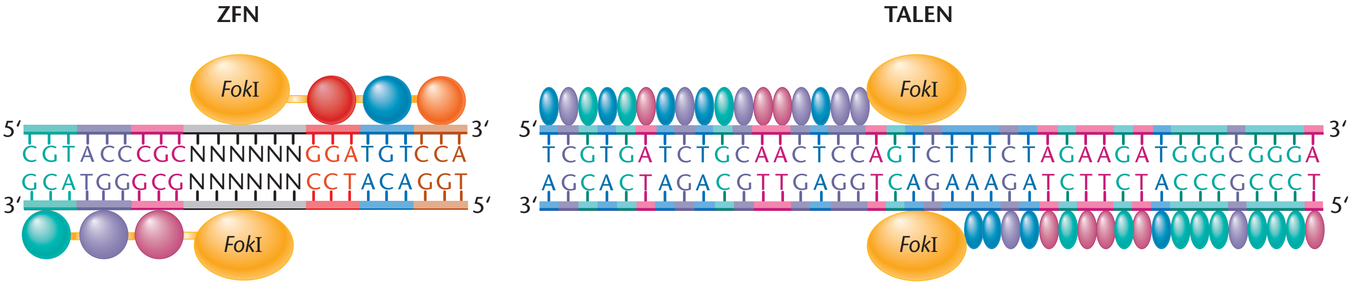


FIGURE ST 5.6 Zinc-finger nucleases and TALENs bind and cut DNA at specific sequences.

of six patients treated, immune-cell counts rose substantially and viral loads also decreased following the therapy. To date, now more than 90 people have been treated by this approach. What percentage of immune cells would have to be treated this way to significantly inhibit spread of the virus is still not known.

Recently, researchers working with human cells used TALENs to remove defective copies of the *COL7A1* gene, which causes recessive dystrophic epidermolysis bullosa (RDEB), an incurable and often fatal disease that presents as excessive blistering of the skin, pain, and severely debilitating skin damage. Researchers at the University of Minnesota used a TALEN to cut DNA near a mutation in the *COL7A1* gene in skin cells taken from a patient with RDEB. These cells were then converted into a type of stem cell called *induced pluripotent stem cells (iPSCs)*. The iPSCs were treated with therapeutic copies of the *COL7A1* gene and then differentiated into skin cells that expressed the correct protein. This is a promising result, and researchers now plan to transplant these skin cells into patients in an attempt to cure them of RDEB. Another group has recently taken a similar approach using TALENs to repair cultured cells in order to correct the mutation in Duchenne muscular dystrophy (DMD). Researchers are optimistic that this approach can soon be adapted to treat patients.

CRISPR-Cas Method Revolutionizes Gene Editing Applications and Renews Optimism in Gene Therapy

No method has created more excitement than the gene editing technique known as **CRISPR-Cas** (clustered regularly interspaced short palindrome repeats—CRISPR-associated proteins). We discuss CRISPR-Cas in much greater detail in Special Topics Chapter 1—CRISPR-Cas and Genome Editing. Identified in bacterial cells, the CRISPR-Cas system functions to provide bacteria and archaea immunity against invading bacteriophages and foreign plasmids. First introduced in 2013, a CRISPR-Cas craze unfolded that has revolutionized genome-engineering applications including gene editing for gene therapy. Because CRISPR-Cas works in bacteria, animal, and plant cells, the method offers diverse applications for genetic engineering by targeted gene editing.

CRISPR-Cas is based on delivering a single-stranded guide RNA sequence (sgRNA) that is complementary to the target gene sequence in the genome and attached to an endonuclease. One commonly used nuclease is called Cas9 (**Figure ST 5.7**). Compared to TALEN approaches, sgRNAs are relatively easy to design and synthesize. At the same time as the sgRNA sequence is delivered, a DNA donor template strand coding for a replacement sequence is delivered. The sgRNA-Cas9 complex binds to the target DNA sequence, and Cas9 generates a blunt, double-stranded break in the DNA. CRISPR-Cas recognition of DNA cleavage sites is determined by RNA-DNA base pairing and a protospacer-adjacent motif (PAM), a three-nucleotide sequence adjacent to the complementary sequence. As cells repair the DNA damage caused by Cas9, repair enzymes incorporate donor template DNA into the genome at the CRISPR-Cas site, thus replacing the target DNA sequence.

CRISPR-Cas is much easier to use than ZFNs or TALENs. Part of the power of CRISPR-Cas is that editing can be done directly in a living, adult organism, in a fairly easy, accurate, and efficient manner. Within months of the technique

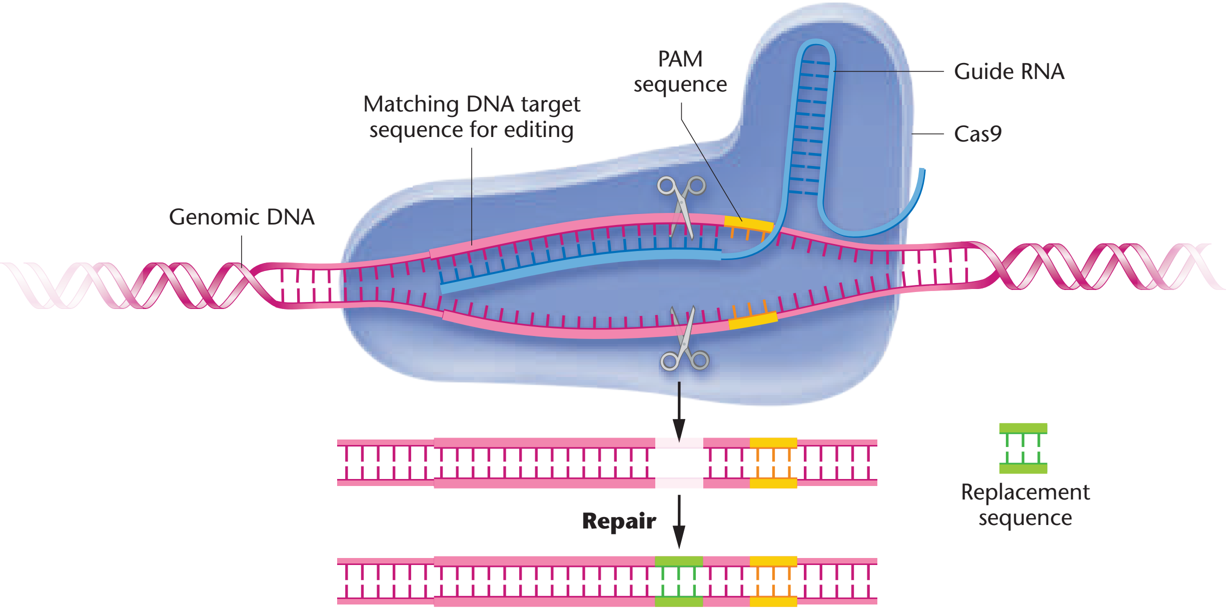


FIGURE ST 5.7 The CRISPR-Cas system allows for gene editing by targeting specific sequences in the genome.

being widely available, researchers around the world used CRISPR-Cas to target specific genes in human cells, mice, rats, bacteria, fruit flies, yeast, zebrafish, and dozens of other organisms. In one of the first reported applications of CRISPR-Cas for gene therapy, a team from the Massachusetts Institute of Technology (MIT) cured mice of a rare liver disorder, type I tyrosinemia, through gene editing. In tyrosinemia, a condition affecting about 1 in 100,000 people, mutation of the *FUH* gene encoding the enzyme fumarylacetoacetase prevents breakdown of the amino acid tyrosine. After an *in vivo* approach with a one-time treatment, roughly 1 in 250 liver cells accepted the CRISPR-Cas–delivered replacement of the mutant gene with a normal copy of the gene. But about 1 month later these cells proliferated and replaced diseased cells, taking over about one-third of the liver, which was sufficient to allow mice to metabolize tyrosine and show no effects of disease. Mice were subsequently taken off a low-protein diet and a drug normally used to disrupt tyrosine production.

In Special Topics Chapter 3 Precision Medicine, we mentioned a promising way to treat cancer called **immunotherapy**. This approach harnesses a patient's own immune system to kill tumors, and some have brought remarkable therapeutic effects in clinical trials (see Figures ST 3–xxx). You were introduced to *engineered or adopted T-cell* methods in which a patient's T cells are genetically engineered *ex vivo* before being returned to the body. The principle behind engineered T-cell therapies is to create T cells that can find and target tumor cells for destruction. Recall that two strategies for immunotherapy are to create recombinant **T-cell receptors (TCRs)** that specifically recognize antigens on or within cancer cells or **chimeric antigen receptor (CAR)–T cells** engineered to express receptors that can directly recognize antigens on the surface of the tumor cell without requiring T-cell activation by antigen-presenting cells. Immunotherapy has shown great promise for the treatment of certain forms of leukemia, and its applications are one area where the ease of gene editing by CRISPR-Cas has had an immediate, positive impact.

In 2016 in the United States, an advisory group from the NIH approved trials for CRISPR-Cas9 editing of T cells from patients with melanoma, multiple myeloma, and other cancers not responding to traditional therapies. In August 2017, after a panel of physicians and researchers who advise the FDA unanimously recommended approval of Novartis' CTL019 CAR-T treatment for children and young adults with B-cell acute lymphoblastic leukemia (ALL), the FDA approved CTL019 as the first CAR-T gene therapy. ALL is the most common childhood cancer in the U.S. and patients who relapse or fail to respond to chemotherapy have a low survival rate. But over >80% of 68 patients treated with CTL109 went into remission almost immediately after treatment began and most remained cancer-free six months after treatment. CTL019 (brand name, Kymriah™) is approved for ALL patients up to 25 years of age. Kymriah also made headlines for its sticker price of approximately $475,000 for a single treatment and estimates that the total cost of care with this drug could exceed $1.5M.

A team from the University of Pennsylvania; the University of California, San Francisco; and the MD Anderson Cancer Center at the University of Texas proposed to treat patient T cells by editing two genes to help these cells target destruction of tumor cells and to avoid nontumor cells. The trial will also edit a gene called *program cell death 1 (PD-1)*, which expresses a protein on the surface of T cells that can often be neutralized by cancer cells to minimize immune responses and ward off T-cell destruction of a tumor. The edited T cells with the mutant *PD-1* gene are expected to be able to recognize and attack lung tumor cells. This trial was scheduled to start in late 2017 in the United States.

In October 2016, a team from China reported the first CRISPR-Cas trial of human patients, designed to treat an aggressive form of cancer called metastatic non–small-cell lung cancer. An *ex vivo* approach was used to edit *PD-1* in T cells, which were then reintroduced into patients with the hope that these cells would target tumors in the lung for destruction without being disabled by the cancer cells. The main purpose of this initial trial is to determine whether this approach is safe. The success of this trial had not been reported at the time this edition of *Concepts of Genetics* was published. Several more gene-editing trials are scheduled to begin at different centers around the world targeting kidney, bladder, and prostate cancer, among others.

Additional headline-grabbing examples of successful CRISPR-Cas applications in mice and humans that highlight the potential of this approach for gene therapy include:

- AAV delivery of CRISPR-Cas9 to remove a defective exon from the *Dmd* gene in a mouse model (*mdx* mice) of DMD significantly restored muscle function in treated mice. It has been estimated that this approach has the potential to cure approximately 80 percent of human cases of DMD.
- After successful trials in mice, *in vitro* repair of the human β-globin (*HBB*) gene in HSCs was used to treat sickle-cell disease and β-thalassemia in humans. It is expected that these preclinical studies will soon lead to the delivery of edited HSCs in humans.
- CRISPR-Cas9 targeting and replacement of the defective clotting factor IX gene in liver cells was used to cure mice of hemophilia B.
- CRISRP-Cas9 editing to disable the *CCR5* gene (which we discussed as a ZFN treatment) to block HIV infection is another approach used that combines immunotherapy (CAR–T-cell therapy) and gene editing.

- A CRISPR-Cas approach to edit the *CEP290* gene to treat LCA type 10 (recall that this is a form of blindness) is currently in clinical trials.
- CRISPR-Cas editing of mutations involved in genetic forms of hearing loss in mice have shown early potential.

Finally, the use of CRISPR-Cas to edit the human germ line (sperm and egg cells) and human embryos has been one of the most controversial potential applications of this technology, although similar concerns existed when ZFNs and TALENs were first being used. In 2015, a team of Chinese scientists reported using CRISPR-Cas9 to edit the *HBB* gene in 86 human embryos donated previously for *in vitro* fertilization. Two days after CRISPR-Cas treatment, 71 embryos had survived, but only 4 of them carried the intended change to the *HBB* gene. Unexpectedly, many other embryos had acquired mutations in genes other than *HBB* as a result of the treatment. From this the Chinese research team concluded that gene editing technology is not sufficiently developed for use in embryos. In 2016, the second published report of gene editing of human embryos, also from a team in China, described the use of CRISPR-Cas9 to introduce a mutation in the *CCR5* gene to confer resistance to HIV infection. In this study, 4 of 26 embryos were successfully edited, but others contained undesirable mutations as a result of the treatment. While this demonstrated some proof of concept for creating HIV-resistant embryos, like the 2015 work, it also clearly showed that gene editing of embryos is neither precise nor safe at this point.

These and other studies have stimulated significant ethical discussions about genetic engineering of embryos (which we discuss further in the final section of this chapter). Currently about two dozen countries (including the United States and the United Kingdom) have a strict ban on human germ-line and embryo modification, but more permissive regulations exist in other countries. This continues to prompt widespread concern from scientists, ethicists, politicians, and patients. Countries such as China, India, and Japan have guidelines, not true legislation, banning genetic modifications, but these are generally considered unenforceable or are largely ignored.

Also, as is the case with other gene therapy approaches, there is concern about CRISPR-Cas creating mutations at nontarget locations in the genome. But so far, CRISPR-Cas is clearly the most promising tool for gene editing and gene therapy that has ever been developed, and in a short time, the pace of progress with this technique in animals and humans has been remarkable. Stay tuned!

RNA-Based Therapeutics

Over the past decade, RNA-based therapeutics such as antisense RNAs, RNA interference, and microRNAs for gene therapy have received a great deal of attention. This is partly because these methods can be designed to be highly specific for a target RNA of interest to block or upregulate gene expression and are versatile because they can also be used to alter mRNA splicing, to target noncoding RNAs, and to express an exogenous RNA among other examples.

Attempts have been made to use **antisense oligonucleotides** to inhibit translation of mRNAs from defective genes (see Figure ST 5.8), thus blocking or "silencing" gene expression, but this approach to gene therapy has generally not yet proven to be reliable. The emergence of **RNA interference (RNAi)** as a powerful gene-silencing tool *in vitro* for research has reinvigorated interest in gene therapy approaches by gene silencing. As you learned earlier in the text (see Chapter 18), RNAi is a form of gene-expression regulation. In animals, short, double-stranded RNA molecules are delivered into cells where the enzyme Dicer chops them into 21- to 25-nt-long pieces called **small interfering RNAs (siRNAs).** siRNAs then join with an enyzme complex called the **RNA-inducing silencing complex (RISC),** which shuttles the siRNAs to their target mRNA, where they bind by complementary base pairing. The RISC complex can block siRNA-bound mRNAs from being translated into protein or can lead to degradation of siRNA-bound mRNAs so that they cannot be translated into protein (Figure ST 5.8).

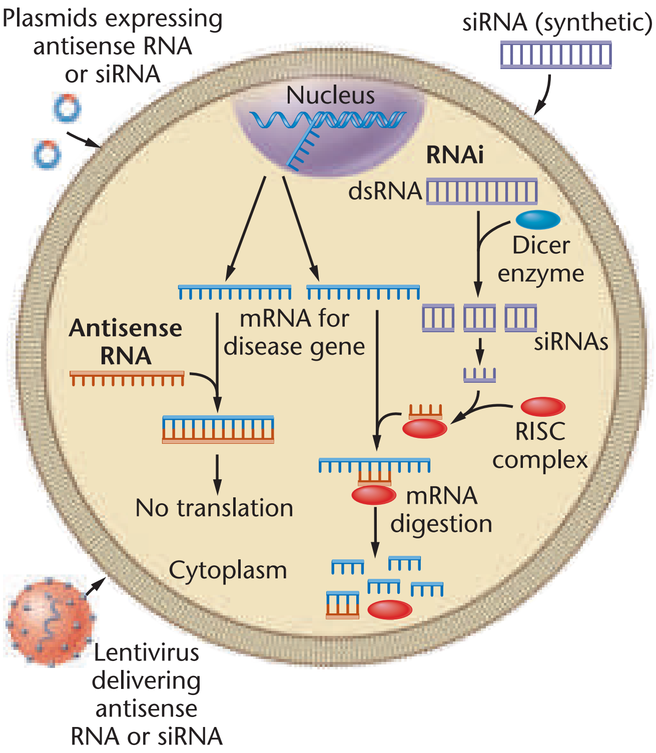


FIGURE ST 5.8 Antisense RNA and RNA interference (RNAi) approaches to silence genes for gene therapy. Antisense RNA technology and RNAi are two ways to silence gene expression and turn off disease genes.

A main challenge to RNAi-based therapeutics so far has been *in vivo* delivery of double-stranded RNA or siRNA. RNAs degrade quickly in the body. It is also hard to get RNA to cross lipid bilayers to penetrate cells in the target tissue. And how does one deliver RNA-based therapies to cancer cells but not to noncancerous, healthy cells? Two common delivery approaches are to inject the siRNA directly or to deliver them via a DNA plasmid vector that is taken in by cells and transcribed to make double-stranded RNA which Dicer can cleave into siRNAs. Lentivirus, liposome, and attachment of siRNAs to cholesterol and fatty acids are other approaches being used to deliver siRNAs (**Figure ST 5.8**). When delivered in liposomes or attached to lipids, siRNAs are taken into the cell by endocytosis, but because of their charge, another challenge is getting therapeutic RNA out of the endosome and into the cytoplasm.

The same approaches used to deliver antisense RNAs and siRNAs can also be used to deliver vectors encoding **microRNAs (miRNAs)** or miRNAs themselves. (Recall that we discussed miRNAs as in Chapter 17.) In recent years a tremendous body of research literature has developed on the roles of miRNAs in silencing gene expression naturally in cells. The application of miRNAs for gene therapy is only in initial stages of development.

More than a dozen clinical trials involving RNAi are under way in the United States. Several RNAi clinical trials to treat blindness are showing promising results. One RNAi strategy to treat a form of blindness called macular degeneration targets a gene called *VEGF.* The VEGF protein promotes blood vessel growth. Overexpression of this gene, causing excessive production of blood vessels in the retina, leads to impaired vision and eventually blindness. Many expect that this disease will soon become the first condition to receive approval for treatment by RNAi therapy. Other disease candidates for treatment by RNAi include several different cancers, diabetes, liver diseases, multiple sclerosis, and arthritis.

But many are predicting that RNA-based therapies will become much more successful and more widely adopted within the next decade. Antisense RNA is the oldest RNA-based therapeutic approach, but it initially did not live up to expectations. In the mid-2000s, many companies dropped this approach to gene therapy because of problems associated with delivering antisense RNA oligonucleotides across cell membranes and keeping them from being degraded while in the circulatory system. However, recent advances in RNA oligonucleotide chemistry have helped overcome these hurdles, and hundreds of clinical trials involving antisense RNAs are in the planning stages.

By late 2016, within a six-month span, antisense oligonucleotide trials were approved by the FDA for familial cholesterolemia, Duchenne muscular dystrophy, and spinal muscular atrophy (SMA). For example, the antisense oligonucleotide called Spinraza® (nusinersen), produced by Ionis Pharmaceuticals of Carlsbad, California, was approved as the first treatment for SMA. Affecting 1 in 10,000–12,000 children born, this disease is characterized by the loss of motor neurons in the spinal cord resulting in progressive muscle weakness and is a leading genetic cause of death for infants. Patients with SMA have a mutation in the *SMN1* gene that prevents production of the functional SMN protein required for normal motor neuron development. Spinraza is delivered into the cerebrospinal fluid. It is an 18-nt antisense oligonucleotide that targets *SMN2*, a homolog of *SMN1*. *SMN2* is normally mis-spliced at exon 7 to produce a truncated, largely nonfunctional SMN protein.

Spinraza binds to pre-mRNA of the *SMN2* gene, altering splicing to include exon 7 in the mature transcript leading to translation of a functional copy of the SMN protein. The drug showed such promise in two trials that the trials were halted early and considered successful because it was deemed unethical to continue to deny the drug to SMN-affected children in placebo groups. This antisense approach to alter mRNA splicing for gene therapy has also generated excitement because it has potential for treating Huntington disease, ALS, and other neurological conditions.

ST 5.7 Future Challenges and Ethical Issues

Despite the progress that we have noted thus far, many questions remain to be answered before widespread application of gene therapy for the treatment of genetic disorders becomes routine:

- What is the proper route for gene delivery in different kinds of disorders? For example, what is the best way to treat brain or muscle tissues?
- What percentage of cells in an organ or a tissue need to express a therapeutic gene to alleviate the effects of a genetic disorder?
- What amount of a therapeutic gene product must be produced to provide lasting improvement of the condition, and how can sufficient production be ensured? Currently, many approaches provide only short-lived delivery of the therapeutic gene and its protein.
- Will it be possible to use gene therapy to treat diseases that involve multiple genes?
- Can expression or the timing of expression of therapeutic genes be controlled in a patient so that genes can be turned on or off at a particular time or as necessary?

- Will gene editing approaches become more widely used for gene therapy trials?
- Will gene editing emerge as the safest and most reliable method of gene therapy, rendering other approaches obsolete, or will a combination of approaches (vector and nonvector delivery, RNA-based therapeutics, and gene editing) be necessary depending on the genetic condition being treated?

For many people, the question remains whether gene therapy can ever recover from past setbacks and fulfill its promise as a cure for genetic diseases. Clinical trials for any new therapy are potentially dangerous, and often, animal studies will not accurately reflect the reaction of individual humans to the methodology leading to the delivery of new genes. However, as the history of similar struggles encountered with life-saving developments such as the use of antibiotics and organ transplants, there will be setbacks and even tragedies, but step by small step, we will move toward a technology that could—someday—provide reliable and safe treatment for severe genetic diseases.

Ethical Concerns Surrounding Gene Therapy

Gene therapy raises several ethical concerns, and many forms of gene therapy are sources of intense debate. At present, in the United States, all gene therapy trials are restricted to using somatic cells as targets for gene transfer. This form of gene therapy is called **somatic gene therapy;** only one individual is affected, and the therapy is done with the permission and informed consent of the patient or family.

Two other forms of gene therapy have not been approved, primarily because of the unresolved ethical issues surrounding them. The first is called **germ-line therapy,** whereby germ cells (the cells that give rise to the gametes—i.e., sperm and eggs) or mature gametes are used as targets for gene transfer or gene editing. In this approach, the transferred or edited gene is incorporated into all the future cells of the body, including the germ cells. As a result, individuals in future generations will also be affected, without their consent. Recently, ethical discussions about germ-line therapy have been accelerated by the prospect of using CRISPR-Cas for gene editing of germ cells and embryos. A report from the U.S. National Academy of Sciences and the National Academy of Medicine is recommending that germ-line therapy trials only be considered for serious conditions for which there is no reasonable alternative treatment option, and where both the risk-benefit options and broad oversight are available. Is this kind of procedure ethical? Do we have the right to make this decision for future generations? Thus far, the concerns have outweighed the potential benefits, and such research is prohibited.

Box 3 mentions gene doping, which is also an example of **enhancement gene therapy,** whereby people may be "enhanced" for some desired trait. This is another unapproved form of gene therapy—which is extremely controversial and is strongly opposed by many people. Should genetic technology be used to enhance human potential? For example, should it be permissible to use gene therapy to increase height, enhance athletic ability, or extend intellectual potential? Presently, the consensus is that enhancement therapy, like germ-line therapy, is an unacceptable use of gene therapy. However, there is an ongoing debate, and many issues are still unresolved.

Gene therapy is currently a fairly expensive treatment. For rare conditions, the fewer the people treated, the more expensive the treatment will be. But what is the right price for a cure? It remains to be seen how health-care insurance providers will view gene therapy. But if gene therapy treatments provide a health-care option that drastically

BOX 3
Gene Doping for Athletic Performance?

Gene therapy is intended to provide treatments or cures for genetic disease. But should it also apply for those seeking genetic enhancements to improve athletic performance? As athletes seek a competitive edge, will gene therapy as a form of "gene doping" to improve performance be far behind?

We already know that in animal models enhanced muscle function can be achieved by gene addition. For example, adding copies of the insulinlike growth factor (*IGF-1*) gene to mice improves aspects of muscle function. The kidney hormone erythropoietin (EPO) increases red blood cell production, which leads to a higher oxygen content of the blood and thus improved endurance. Synthetic forms of EPO are banned in Olympic athletes. Several groups have proposed using gene therapy to deliver the *EPO* gene into athletes "naturally."

Since 2004 the World Anti-Doping Agency (WADA) has included gene doping through gene therapy as a prohibited method in sanctioned competitions. However, methods to detect this are not well established. If techniques for gene therapy become more routine, many feel it is simply a matter of time before gene doping will be the next generation of performance-enhancement treatments. Obviously, many legal and ethical questions will arise if this becomes a reality.

SPECIAL TOPIC 5

improves the quality of life for patients for whom there are few other options, it is likely that insurance companies will reimburse patients for treatment costs.

Finally, *whom* to treat by gene therapy is yet another ethically provocative consideration. In the Jesse Gelsinger case mentioned earlier, the symptoms of his OTC deficiency were minimized by a low-protein diet and drug treatments. Whether it was necessary to treat Jesse by gene therapy is a question that has been widely debated.

Jesse Gelsinger volunteered for the study to test the safety of the treatment for those with more severe disease. If a benefit was shown, it would have relieved him of an intense treatment regimen. Whether he should have been selected for the safety study is, of course, a matter to be debated. His tragic death due to unforeseen complications could not have been predicted at the time.

Selected Readings and Resources

Journal Articles

Baltimore, D., et al. (2015). A prudent path forward for genomic engineering and germline gene modification. *Science* 348:36–38.

Dever, D. P., et al. (2016). CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. *Nature* 539:384–389.

Dowdy, S. F. (2017). Overcoming cellular barriers for RNA therapeutics. *Nature Biotech*. 35:222–229.

Gaj, T. et al. (2016). Genome engineering using adeno-associated virus: basic and clinical research applications. *Molecular Therapy*, 24: 458-464.

Hirsch, T., et al. (2017). Regeneration of the entire human epidermis using transgenic stem cells. *Nature*, 551: 327-332.

Ma, H. et al. (2017). Correction of a pathogenic gene mutation in human embryos. *Nature*, 548:413-419.

Maeder, M.L., and Gersbach, C.A. (2016). Genome-editing technologies for gene and cell therapy. *Molecular Therapy*, 24: 430-446.

Nelson C. E., et al. (2016). In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science* 351:403–405.

Orkin, S. H., and Reilly, P. (2016). Paying for future success in gene therapy. *Science* 352:1059–1061.

Prakash, V. et al. (2016). Current progress in therapeutic gene editing for monogenic diseases. *Molecular Therapy*, 24: 465-474.

Sadelain, M., Rivière, I., and Riddell, (2017). S. Therapeutic T cell Engineering. *Nature*, 545:423-431.

Scoles, D.R., et al. (2017). Antisense oligonucleotide therapy for spinocerebellar ataxia type 2. Nature 544: 362-371.

Review Questions

1. What is gene therapy?
2. Compare and contrast *ex vivo* and *in vivo* gene therapy as approaches for delivering therapeutic genes.
3. When treating a person by gene therapy, is it necessary that the therapeutic gene becomes part of a chromosome (integration) when inserted into cells? Explain your answer.
4. Describe two ways that therapeutic genes can be delivered into cells.
5. Explain how viral vectors can be used for gene therapy, and provide two examples of commonly used viral vectors. What are some of the major challenges that must be overcome to develop safer and more effective viral vectors for gene therapy?
6. During the first successful gene therapy trial in which Ashanti DeSilva was treated for SCID, did the therapeutic gene delivered to Ashanti replace the defective copy of the ADA gene? Why were white blood cells chosen as the targets for the therapeutic gene?
7. Explain an example of a successful gene therapy trial. In your answer be sure to consider: a description of the disease condition that was treated, the mutation or disease gene affected, the therapeutic gene delivered, and the method of delivery use for the therapy.
8. What is gene editing, and how does this approach differ from traditional gene therapy approaches?
9. What are some of the reasons why the development of CRISPR-Cas may be the technical breakthrough that will make gene therapy a safe and more common treatment for many genetic conditions?
10. How do ZFNs work?
11. Describe two gene-silencing techniques, and explain how they may be used for gene therapy.

Discussion Questions

1. Discuss the challenges scientists face in making gene therapy a safe, reliable, and effective technique for treating human disease conditions.
2. Who should be treated by gene therapy? What criteria are used to determine if a person is a candidate for gene therapy? Should gene therapy be used for cosmetic purposes or to improve athletic performance?
3. The lifetime costs for treatment of conditions such as hemophilia A and sickle-cell disease can be several million dollars. Many immunotherapies and Glybera (see Box 2) treatment cost over $1 million per patient. What is the appropriate way to determine the price for a gene therapy treatment? Who should pay? The patient? Insurance?
4. Should CRISPR-Cas or other techniques be used for editing human germ cells and/or embryos? What safety concerns might need to be addressed to support genome editing in humans?
5. Describe future challenges and ethical issues associated with gene therapy.

Advances in Neurogenetics: The Study of Huntington Disease

As the result of groundbreaking advances in molecular genetics and genomics made since the 1970s, new fields in genetics and related disciplines have emerged. One new field is **neurogenetics**—the study of the genetic basis of normal and abnormal functioning of the nervous system, with emphasis on brain functions. In addition to extending our understanding of how the nervous system and the brain work, research in this field is focused on the genes associated with neurodegenerative disorders, with the ultimate goal of developing effective therapies to combat these devastating conditions. Of the many such diseases, including Alzheimer disease, Parkinson disease, and amyotrophic lateral sclerosis (ALS), **Huntington disease (HD)** stands out as a model for the genetic investigation of neurodegenerative disorders. Not only is it monogenic and 100 percent penetrant, but nearly all analytical approaches developed over several decades in molecular genetics have been successfully applied to the study of HD, validating its significance as a model for these diseases.

> "Driving with my father through a wooded road leading from Easthampton to Amagansett, we suddenly came upon two women, mother and daughter, both bowing, twisting, grimacing. I stared in wonderment, almost in fear. What could it mean?"

HD is inherited as an autosomal dominant disorder characterized by adult onset of defined and progressive behavioral changes, including uncontrolled movements (chorea), cognitive decline, and psychiatric disturbances, with death occurring within 10–15 years after symptoms appear. HD was one of the first examples of complete dominance in human inheritance, with no differences in phenotypes between homozygotes and heterozygotes. While there is a juvenile form of HD, in the vast majority of cases, symptoms do not develop until about age 45. Overall, HD currently affects about 25,000–30,000 people in North America.

The disease is named after George Huntington, a nineteenth-century physician. He was not the first to describe the disorder, but his account was so comprehensive and detailed (see Box 1) that the disease eventually took on his name. Further, his observation of transgenerational cases in several families precisely matched an autosomal dominant pattern of inheritance. Shortly after the rediscovery of Mendel's work in the early twentieth century, pedigree analysis confirmed that HD is inherited as an autosomal dominant disorder.

We will begin our consideration of Huntington disease by discussing the successful efforts to map, isolate, and clone the HD gene. We will discuss our ability to test those at risk for the disorder and the ethical issues that diagnosis of such a genetic disorder poses. We will then turn our attention to what we know about the molecular and cellular mechanisms associated with the disorder, particularly those discovered during the study of transgenic model systems. We will also consider how this information is being used to develop a range of therapies, and finally, we will discuss the relationship between HD and other neurodegenerative disorders.

BOX 1

George Huntington and His Namesake Disease

George Huntington first encountered the disease that would later bear his name as an 8-year-old boy riding in a horse-drawn carriage with his father, a local physician from Long Island, New York:

> "Driving with my father through a wooded road leading from Easthampton to Amagansett, we suddenly came upon two women, mother and daughter, both bowing, twisting, grimacing. I stared in wonderment, almost in fear. What could it mean? My father paused to speak with them, and we passed on. . . . From this point, my interest in the disease has never wholly ceased."

Later in 1872 as a physician at age 22, he published a paper providing a definitive description of this disorder, which at the time was known to affect the nervous system, causing uncontrollable twitches and limb movements called *chorea* (a word that means dance in ancient Greek). After his description, this disorder came to be known as Huntington's chorea and is now called Huntington disease. From thorough observations of patients and their families, Huntington arrived at several important conclusions:

1. The disorder is hereditary. He accurately described a pattern of autosomal dominant inheritance, several decades before Mendel's paper was brought to wider attention:

   > "When either or both the parents have shown manifestations of the disease, and more especially when these manifestations have been of a serious nature, one or more of the offspring almost invariably suffer from the disease, if they live to adult age. But if by any chance these children go through life without it, the thread is broken and the grandchildren and great-grandchildren of the original shakers may rest assured that they are free from the disease."

2. Progressive cognitive deficits and dementia are an important part of the disease:

   > "As the disease progresses, the mind becomes more or less impaired, in many amounting to insanity, while in others mind and body both gradually fail until death relieves them of their sufferings. "

3. The disease has an adult onset and is incurable:

   > "Its third peculiarity is its coming on, at least as a grave disease, only in adult life. I do not know of a single case that has shown any marked signs of chorea before the age of thirty or forty years, while those who pass the fortieth year without symptoms of the disease are seldom attacked." . . . "I have never known a recovery or even an amelioration of symptoms in this form of chorea; when once it begins it clings to the bitter end. No treatment seems to be of any avail, and indeed nowadays its end is so well-known to the sufferer and his friends, that medical advice is seldom sought. It seems at least to be one of the incurables."

Unfortunately, what Huntington wrote about this disease in the nineteenth century remains true today. In spite of the many advances made since the gene for Huntington disease was identified in 1993, there is no treatment yet available to slow or reverse the inevitable and relentless progression of this terminal disease.

ST 6.1 The Search for the Huntington Gene

Mapping the gene for Huntington disease was one of the first attempts to employ a method from a landmark 1980 paper by Botstein, White, and Davis[*] in which the authors proposed that DNA sequence variations in humans could be detected as differences in the length of DNA fragments produced by cutting DNA with restriction enzymes. These differences, known as restriction fragment length polymorphisms (RFLPs), could be visualized using Southern blots (see Chapter 22 for a discussion of RFLPs, and Chapter 20 for a discussion of Southern blots). The authors estimated that a collection of about 150 RFLPs distributed across the genome could be used with pedigrees to detect linkage anywhere in the genome between an RFLP marker and a disease gene of interest. In practical terms, this meant that it would be possible to map a disease gene with no information about the gene, its gene product, or its function—an approach referred to as reverse genetics.

Finding Linkage between Huntington Disease and an RFLP Marker

In the early 1980s, Huntington disease research was largely driven by the Hereditary Disease Foundation, established by the family of Leonore Wexler, who, along with her three brothers, died of Huntington disease. One daughter, Nancy, after learning about the proposal to map disease genes using DNA markers, used her awareness of a large population affected with Huntington disease in

[*] Key papers in this Special Topics are listed in the end-of-chapter Selected Readings and Resources.

BOX 2

Nancy Wexler and the Venezuelan Pedigree

Nancy Wexler's lifelong involvement with Huntington disease research began when she learned that her mother was diagnosed with this fatal disorder. Soon after, she began working with the Hereditary Disease Foundation established by her father to sponsor workshops and to encourage research on this disorder. Later she was named executive director of a U.S. Congressional Commission for the Control of Huntington Disease to establish funding priorities for research. The Commission formed the Venezuela Working Group in 1979 to study a large cluster of Huntington cases in villages surrounding Lake Maracaibo. Over the next 22 years, teams of workers went to Venezuela twice a year to gather family pedigrees and to collect blood samples from a subset of pedigree members. The pedigree eventually included 18,000 individuals. Just over 4000 cell lines were established from blood samples. DNA from these cell lines was used to locate and identify the gene for Huntington disease.

Venezuela to organize trips to collect pedigree information and to obtain blood samples for DNA linkage studies (see Box 2).

About the same time Nancy Wexler began working on the Venezeulan pedigree, James Gusella began collecting RFLP markers to map the gene for Huntington disease. Instead of waiting to amass a collection of 150–300 markers and then mapping them to individual chromosomes before searching for the gene, Gusella decided to use the 13 markers he had available. Using these markers and DNA from an American HD family, Gusella failed to find clear evidence of linkage to HD. Because these 13 markers were not informative, in the next round of experiments Gusella decided to set aside these markers and use only his most recently developed RFLP marker, called G8. The G8 marker identified four possible patterns of DNA fragments produced by the restriction enzyme *Hind*III. The patterns, called haplotypes, were named A, B, C, and D. In addition, he decided to use DNA from a large Venezuelan HD pedigree instead of DNA from the American HD family.

After analyzing the results from Southern blots of the Venezuelan pedigree, Gusella's team concluded there was linkage between the gene for Huntington disease and haplotype C (**Figure ST 6.1**). For confirmation, they sent their results to Michael Conneally, an expert in linkage analysis. Conneally verified that the evidence for linkage between HD and haplotype C was overwhelming, and that Gusella and his colleagues had discovered linkage to HD.

Once the G8 probe was linked to the HD gene, the next task was to determine which human chromosome carried the G8 marker and the gene for Huntington disease.

Assigning the HD Gene to Chromosome 4

A collection of somatic cell hybrids (see Chapter 5 for a description of this method) can be used to map DNA markers or genes to specific chromosomes. In this case, a panel of 18 mouse–human somatic cell hybrid cell lines, each of which contained a unique combination of human chromosomes, was used for mapping the G8 probe. On Southern blots from these hybrid cells digested with *Hind*III, G8 fragments were seen in all cells carrying human chromosome 4 and never seen when chromosome 4 was absent.

These results, reported by Gusella, Wexler, and colleagues in 1983, established that the G8 marker and the gene for Huntington disease were both on chromosome 4. This was the first time that an RFLP marker was used to map an autosomal disease gene to a specific chromosome. This discovery launched a whole new branch of genetics, called *positional cloning* (sometimes called reverse genetics). In short order, genes for several other genetic disorders were mapped, revolutionizing the field of human genetics.

Over the next few years, studies utilizing several different methods localized the G8 marker (now renamed as

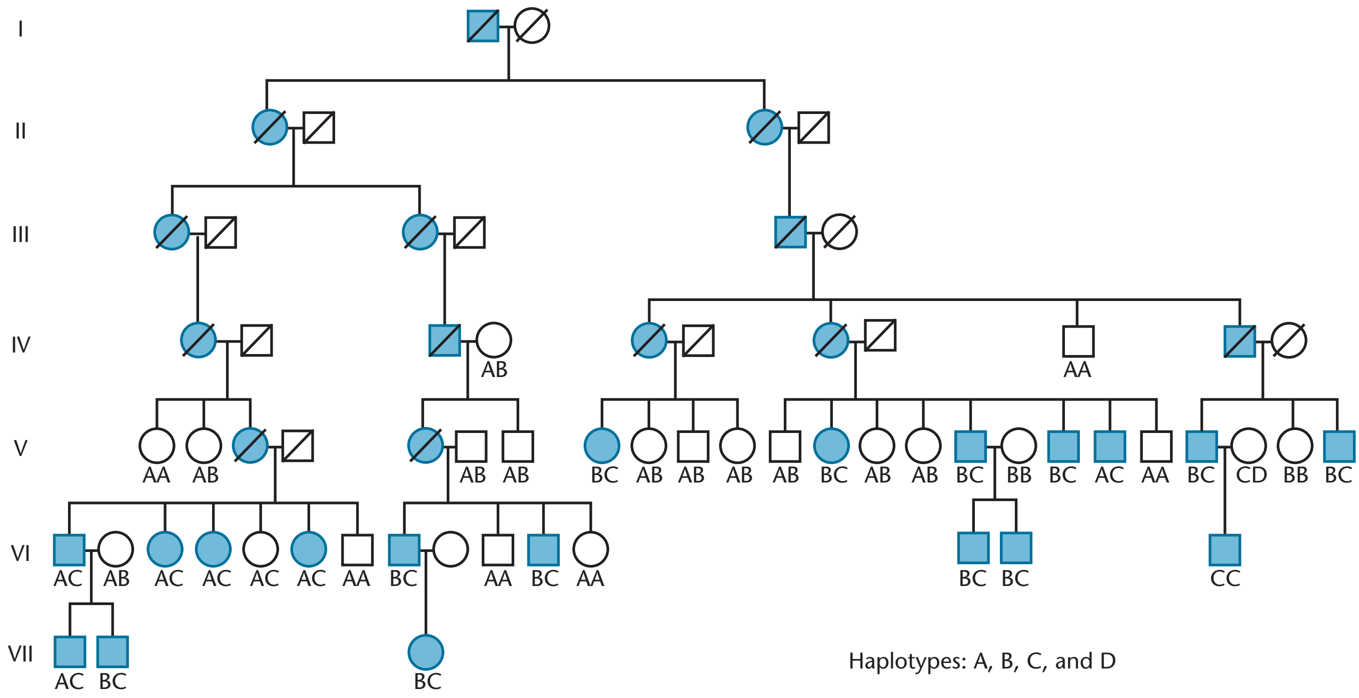

I
II
III
IV
AB
AA
V
AA AB
AB AB
BC AB AB AB AB BC AB AB BC BB BC AC AA BC CD BB BC
VI
AC AB AC AC AC AC AA BC AA BC AA
BC BC
CC
VII
AC BC
BC
Haplotypes: A, B, C, and D

FIGURE ST 6.1 A part of the Venezuelan pedigree used in the search for linkage between RFLP markers and Huntington disease. Filled symbols indicate affected individuals. Deceased individuals are marked by diagonal slashes. In this pedigree, haplotype C of the G8 marker is coinherited with HD in all cases, indicating that the RFLP marker and the mutant HD allele are on the same chromosome.

D4S10) to a region containing a few million base pairs at the tip of the short arm of chromosome 4 (**Figure ST 6.2**). While this narrowed the search considerably, the exact location of the HD gene remained unknown.

The Identification and Cloning of the Huntington Gene

To identify and clone the gene for Huntington disease, researchers formed The Huntington's Disease Collaborative Research Group, consisting of 58 scientists on two continents. In spite of this massive effort, it took 10 years to identify the gene. First, the region most likely to contain the gene was narrowed to a 2.2-Mb region within chromosome band 4p16.3. Analysis of shared haplotypes from individuals with HD further narrowed the region to a 500-kb segment within band 4p16.3. Overlapping clones (contigs) covering this region were constructed and screened by a method called *exon trapping* to extract exons (coding regions) from these clones. These trapped exons were used to screen cDNA libraries to isolate expressed genes that the research team named "interesting transcripts (ITs)." One of the genes identified in this screening, called *IT15*, encoded a previously unknown protein of about 348 kd. Exon 1 of this gene contained a CAG repeat sequence. Populations unaffected with HD carried more than 17 different alleles of this gene, with CAG repeats ranging from 11–34 copies. However, in individuals with HD, CAG repeats were significantly longer, ranging from 42–66 copies. None of the other genes identified in this 500-kb region had any differences between affected and unaffected individuals that would implicate them as the HD gene. Because variations in the size of trinucleotide repeats had previously been identified as the causes of myotonic dystrophy and fragile-X syndrome, researchers proposed that variation in the number of CAG repeats was the cause of HD and that *IT15* encoded the HD gene.

A paper authored by all 58 members of The Huntington's Disease Collaborative Research Group (HDCRG) was published in 1993, ending the decade-long search for the gene, now called *HTT*, and its encoded protein, which they named huntingtin. Subsequent analysis of CAG repeat lengths in populations of unaffected and affected individuals clarified the relationship between repeat length and the onset of HD (see Box 3).

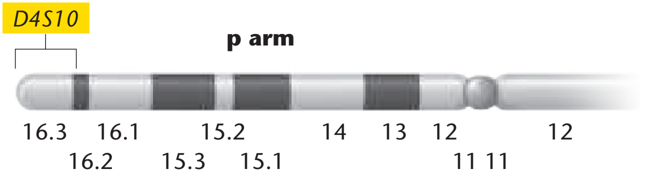


FIGURE ST 6.2 A physical map of the short arm of chromosome 4 showing the location of *D4S10* near the tip of the chromosome in the 4p16.3 region, an area that encompasses about 3 percent of the length of the chromosome.

BOX 3

Genetic Testing for Huntington Disease

International guidelines permit testing for HD from embryonic and fetal stages through childhood and adulthood. Each of these circumstances has an accompanying set of ethical considerations. Guidelines are available from the Web site of the Huntington's Disease Society of America (HDSA).

In all cases, individuals should be made aware of the possible negative psychological and social impact of testing, and genetic counseling and psychological support should be offered before and after testing. In addition, before testing, the guidelines recommend clinical neurological testing for HD symptoms, a psychiatric evaluation, informed consent, and a guarantee of confidentiality for test results.

We will briefly consider the ethical issues in several testing situations. Prenatal testing allows a parent who is at risk for carrying a mutant allele to learn the genetic status of an embryo or fetus. If test results show that the embryo or fetus carries an HD allele, this reveals that at least one parent also carries the mutant allele. Guidelines require that couples where one person is known to be at risk receive genetic counseling about all possible outcomes before conception. Adults who learn they carry a mutant HD allele may express concern that their minor children may also carry this allele and request testing for their child or children. Ethical guidelines consider that the child's rights take precedence over the parent's wishes to know and strongly recommend that minors not be tested, unless there is a clinical indication that a child may have a case of juvenile HD. At age 18, children are free to decide whether or not to be tested.

Presymptomatic testing is done on individuals who are not showing any symptoms of HD but have an affected parent. In this case, the outcome will establish with a high degree of accuracy whether or not the individual will develop HD. Testing of adults with clinical symptoms of HD can rule out other disorders and confirm that a family member does carry the mutant allele. Testing is done by PCR analysis of CAG repeats, sometimes followed by Southern blot analysis. Results are sorted into four risk categories, as described in the following table.

Number of CAG Repeats	Risk Analysis
≤ 26	Normal alleles. No risk.
27–35	Intermediate alleles. No risk. May pose risk for next generation.
36–39	Incomplete penetrance. May or may not cause HD in present and future generations.
≥ 40	Will get HD with 50% chance of transmission to next generation.

Because interpretation and explanation of results can be complex, guidelines suggest that the results be delivered in person at a meeting with a genetic counselor, a clinician, and a psychiatrist or psychologist present.

ST 6.2 The *HTT* Gene and Its Protein Product

Huntington disease is caused by the expansion of a CAG repeat and is one of 14 known trinucleotide repeat disorders. Nine of these, including HD, are caused by the expansion of CAG repeats, each of which codes for the insertion of the amino acid glutamine in the protein product. Thus, these genetic conditions are known as polyglutamine or polyQ disorders (Q is the one letter abbreviation for glutamine). In addition to carrying mutant alleles with expansion of CAG repeats, other polyQ disorders have many symptoms in common with HD, including adult onset, behavioral changes, neurodegeneration, and premature death. The other six trinucleotide repeat disorders do not involve expansion of CAG sequences and contain different trinucleotide repeats, but share crucial molecular and cellular defects with HD.

The *HTT* gene encodes a large protein that is 348–350 kDa. In normal alleles, a region near the 5′ end of the gene contains the 6–35 CAG repeats, encoding a stretch of glutamines in the protein product. Disease-causing mutant alleles contain an expanded number of CAG repeats (>36) that increase the number of glutamine residues in the mutant protein. The normal HTT protein is expressed in most, if not all, cells of the body and is associated with many different cellular compartments and organelles, including the plasma membrane, nucleus, cytoskeleton, cytoplasm, endoplasmic reticulum, Golgi complexes, and mitochondria. In brain cells of the striatum and caudate nucleus, HTT is present at synapses. The HTT protein contains three domains involved in protein–protein interactions and domains consistent with transport of molecules from nucleus to cytoplasm. Posttranslational modifications of HTT confirm its role in facilitating vesicle transport at synaptic junctions. In sum, normal HTT is multifunctional and may undergo conformational changes depending on its location and specific role in cellular processes.

The HD mutation is a gain-of-function mutation. The extended polyQ region of the mutant HTT protein (called

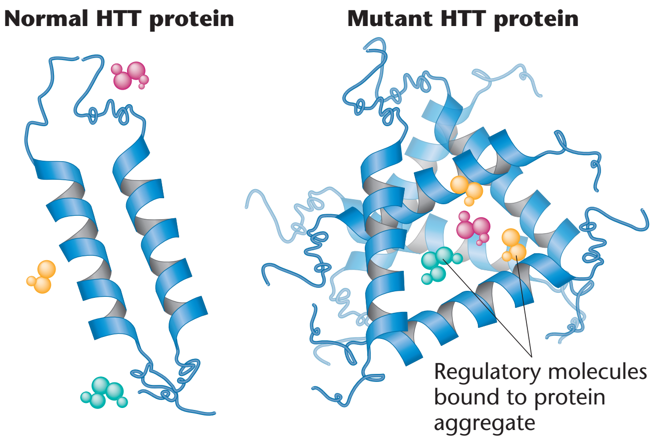


FIGURE ST 6.3 In HD, misfolded mHTT proteins clump together to form aggregates that disrupt cellular functions, partly by binding and sequestering regulatory molecules essential for normal cellular tasks.

mHTT) causes misfolding and the formation of aggregates held together by hydrogen bonds. PolyQ regions of these aggregates bind to and inactivate regulatory molecules (**Figure ST 6.3**), disrupting a number of cellular functions, leading to neurodegeneration. In addition, toxic peptide fragments generated by proteolysis of aggregates are transported into the nucleus where they accumulate and disrupt transcription and nucleocytoplasmic transport. The net result of these cellular changes is a gradually increasing degradation of cellular function that culminates in neurodegeneration and cell death.

ST 6.3 Molecular and Cellular Alterations in Huntington Disease

Although HD is caused by the mutation of a single gene, which was isolated and characterized over 20 years ago, the *mechanisms* by which mutant forms of the HTT protein (mHTT) cause HD are still largely unknown. In the adult brain, one of the major functions of the normal protein appears to be regulation of apoptosis (see Chapter 24 for a discussion of apoptosis). In spite of decades of research employing a wide range of techniques and model systems, the full range of functions carried out by the normal HTT protein and those of the mHTT protein have not been completely elucidated. The straightforward view is that the mutant allele encodes a toxic protein that causes cell death initially in the striatal region of the brain. Increasing loss of cells in the striatum and other regions results in progressive and degenerative changes in muscle coordination and behavior. Death usually occurs 10-15 years after symptoms appear.

Unraveling the functions of the normal and mutant versions of HTT is proving to be extremely difficult because HTT interacts with more than 180 different proteins. Protein—protein network maps indicate that network proteins are involved in many cellular processes including transcription, protein degradation, protein folding, synaptic transmission, and mitochondrial function. Each of these processes malfunctions in the presence of mHTT (**Figure ST 6.4**). In the following sections we will briefly review some of the major pathways affected by expression of the mutant protein.

Transcriptional Disruption

The effects of mHTT on the transcriptome are one of the key molecular events in HD. Expression of mHTT blocks the action of histone acetyl transferases (HATs), causing histone hypoacetylation and formation of heterochromatin, effectively closing off transcription of genes located in affected chromosome regions. In addition, smaller soluble mHTT oligomers interact with components of transcription and obstruct the functions of the promoter-binding transcription factors and the proteins necessary for transcription initiation. The net result is reduced promoter accessibility and transcription initiation across the genome.

Impaired Protein Folding and Degradation

The correct folding of proteins depends on the action of proteins, called chaperones, that mediate folding. In HD, transcriptional disruption inactivates several families of chaperones, which leads to the accumulation of incorrectly folded proteins in the cytoplasm. The result is disruption of normal folding and degradation of cellular proteins.

In normal cells, a small protein, ubiquitin, tags other proteins that are misfolded and directs them to a cellular structure called the proteasome (see Chapter 14 for a discussion of proteasomes) where they are degraded. In brain cells of individuals with HD, ubiquitin binds to aggregated mHTT in the cytoplasm, but protein breakdown by the proteasome system is inhibited by an unknown mechanism. As a result, mHTT aggregates impair cellular functions, triggering apoptosis and cell death. Together, the inhibition of chaperone function and proteasome function causes a collapse of normal protein function and turnover in brain cells of individuals with HD.

Synaptic Dysfunction

In individuals with HD, subtle changes in motor function resulting from synaptic dysfunction can appear decades before the onset of neuronal death. To investigate synaptic defects, researchers constructed transgenic *Drosophila* strains carrying a mutant human HD allele. In transgenic flies, the synaptic vesicles carrying neurotransmitters were much smaller than normal. As a result, synaptic transmission was disrupted, causing behavioral changes in locomotion.

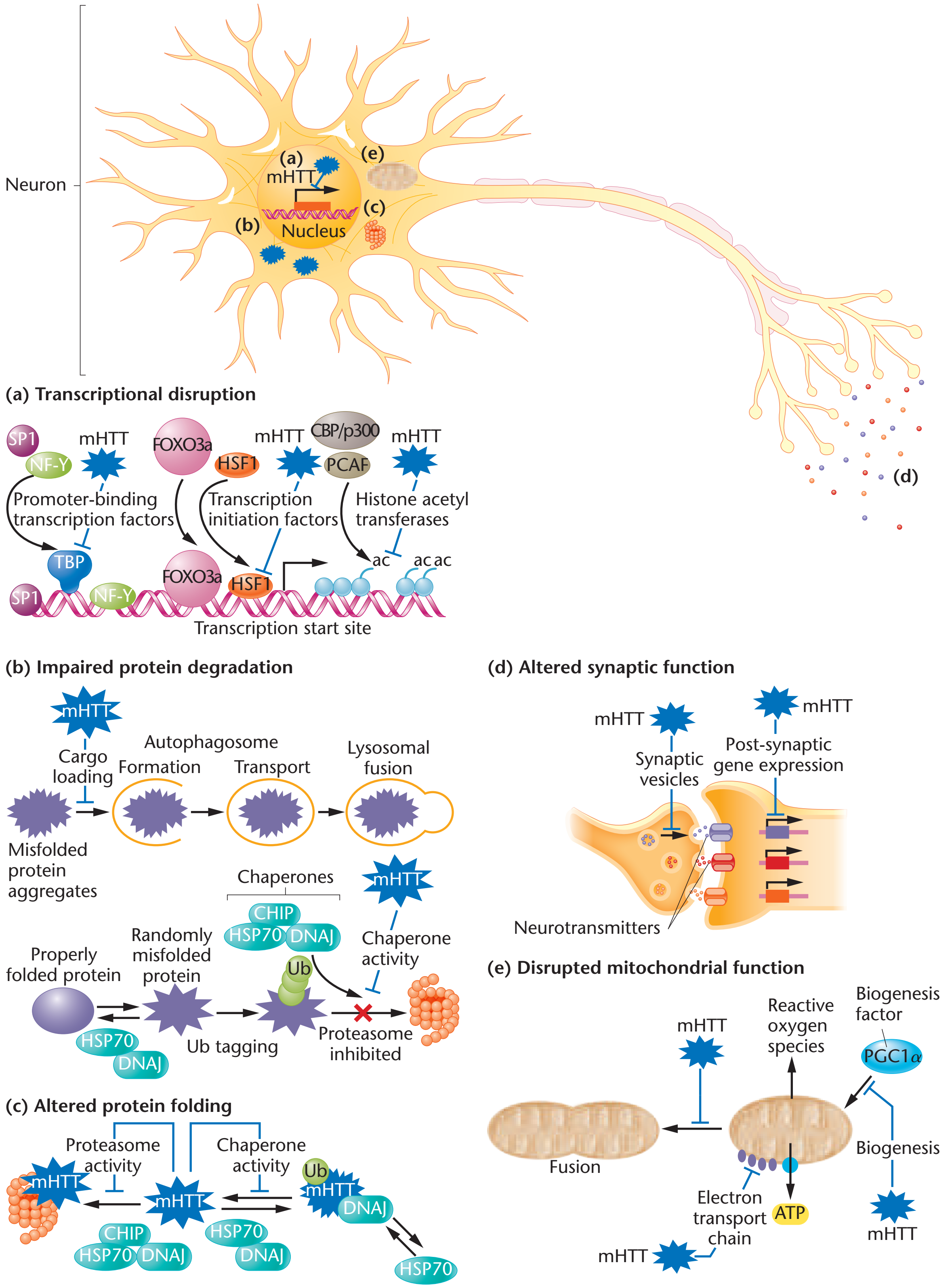


FIGURE ST 6.4 The major pathways of cellular functions disrupted in HD. The letters (a-e) within the neuron correspond to the details of the disrupted processes within the nucleus and cytoplasm shown in parts a-e of the figure. (a) Transcriptional disruption occurs by blocking access of transcription factors and interfering with histone acetylation. (b) Impaired protein degradation is caused by blocking the loading of large protein aggregates into vesicles called autophagosomes for transport to lysosomes and by hindering transport of ubiquitin-tagged proteins to proteasomes. (c) Altered protein folding caused by expanded copies of glutamines that destabilize normal protein conformation and by a lack of chaperones to direct proper refolding along with reduced proteasome activity leads to accumulation of mHTT. (d) Altered synaptic function results from blockage of synaptic vesicle transport by mHTT. (e) Disruption of mitochondrial function by binding of mHTT to the outer membrane triggers a cascade of problems with function, transport, biogenesis, and maintenance.

The Rab family of proteins is involved in the formation, transport, and recycling of synaptic vesicles in neurons. In the presence of mHTT, the activity of one Rab protein, Rab11, is disrupted. Overexpression of Rab11 reversed the deficits in vesicle size, restored normal synaptic transmission of nerve impulses, and reestablished normal locomotion.

Because synaptic dysfunction begins before significant loss of neurons occurs, methods to increase expression of Rab11 may be a useful therapeutic approach to the treatment of HD.

Impaired Mitochondrial Function

mHTT binds to the outer mitochondrial membrane and impairs electron transport, reducing the amount of ATP available to the cell. Disruption of the electron transport chain also increases the levels of reactive oxygen species including free radicals, which cause widespread oxidative damage to cellular structures.

Within neurons, mitochondria migrate to synapses when rates of nerve impulse transmissions increase. Mitochondrial movement is inhibited by aggregation of N-terminal mHTT fragments that physically block migration along microtubules. This reduces the energy available for transmission of nerve impulses at synapses.

In sum, the damage to mitochondria caused by expression of mHTT includes the disruption of ATP production, promotion of oxidative damage within mitochondria and the cytoplasm, lowered synaptic transmission, and reduction of mitochondrial numbers to a level that can no longer support the core activities of cells, which eventually triggers apoptosis and cell death.

Although progress has been made in elucidating the defects that follow the expression and accumulation of mHTT, several questions about the underlying mechanisms of HD remain unanswered. For example, it is not known whether any single disruption of cell function is sufficient to cause neurodegeneration or cell death, or whether one or more of these pathways must interact to bring about these results. To answer these and other questions, researchers turned to the use of transgenic animal models of HD.

ST 6.4 Transgenic Animal Models of Huntington Disease

Shortly after the *HTT* gene was cloned, researchers constructed transgenic model organisms to analyze the disease process at the molecular level.

Animal models of human behavioral disorders present an opportunity to separate behavioral phenotypes into their components. This makes it possible to study the developmental, structural, and functional neuronal mechanisms related to these behaviors that are difficult or impossible to do in humans. In addition, behavior in animal models can be studied in controlled conditions that limit the impact of environmental factors. Although it is possible to construct transgenic models of human HD in a wide range of organisms, including yeast, *C. elegans*, and *Drosophila melanogaster*, the mouse is the most widely used model organism for these studies. Researchers favor mice because humans and mice share about 90 percent of their genes and because a wide range of strains with specific behavioral phenotypes are available.

Using Transgenic Mice to Study Huntington Disease

The first mouse model of HD was constructed using the promoter sequence and first exon of the human mutant *HTT* allele, which contains an expanded CAG repeat. Examination of transgenic mouse brains a year after gene transfer showed abnormalities in the levels of neurotransmitter receptors and the presence of protein aggregates, a significant finding that was later confirmed to exist in the brains of humans with HD.

Soon after, researchers began to examine the relationship between CAG repeat length and disease progression. This work used transgenic mice carrying full-length copies of human *HD* genes with 16, 48, or 89 CAG repeats. Transgenic mice were monitored from birth to death to determine the age of onset and stages of abnormal behavior. Mice carrying 48- or 89-repeat human *HD* genes showed behavioral abnormalities as early as 8 weeks, and by 20 weeks they showed problems with motor coordination.

At various ages, brains of wild-type and transgenic mice carrying mutant alleles with 16, 48, and 89 copies of the CAG repeat were examined for changes in structure. Degenerating neurons and cell loss were evident in mice carrying 48 and 89 repeats, but no changes were seen in brains of wild-type mice or of those carrying a 16-repeat transgene (**Figure ST 6.5**).

There are now more than 20 different mouse models of HD, including mice carrying truncated and full-length copies of the human HD gene, as well as knock-out and knock-in models with modified copies of the mouse homolog to the human *HTT* gene. These mouse models are used to examine changes in molecular and cellular processes or in brain structure that occur before or just after the onset of symptoms, and to develop experimental treatments to slow or reverse cell loss. Transgenic models for HD allow researchers to administer treatment at specific times in disease progression and to evaluate the outcome of treatments in the presymptomatic stages of HD, something that is not possible in humans with HD.

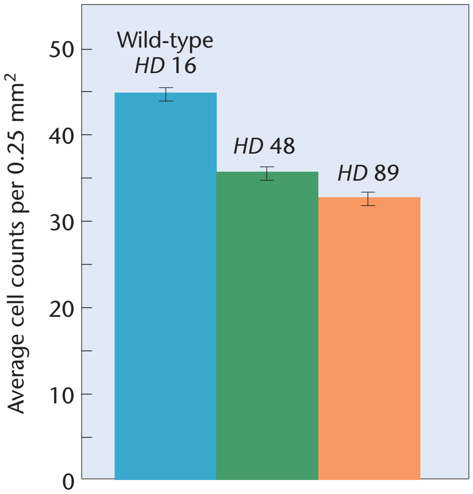


FIGURE ST 6.5 Relative levels of neuronal loss in HD transgenic mice. Cell counts show a significant reduction of neurons in the striatum in the brains of *HD* 48 mutants (middle column) and *HD* 89 mutants (right column). Cell loss in this brain region is also found in humans with HD, making these transgenic mice valuable models to study the course of this disease.

In one important study, A. Yamamoto and colleagues constructed a transgenic mouse with inducible expression of the *HTT* gene. In this case, researchers constructed a human *HTT* exon 1 fragment containing 94 CAG repeats with an adjacent promoter that could be switched off when the antibiotic doxycycline was added to the drinking water. When the gene was switched off shortly after motor symptoms of HD developed, protein aggregates in the brain were rapidly degraded and disappeared, along with the abnormal motor symptoms. This provided the first clue that treatment in the early stages of the disease might be effective in controlling or reversing the symptoms in humans.

Transgenic Sheep as an Animal Model of Huntington Disease

Despite the important discoveries made in mice, the main problem with transgenic mouse models of HD is that mice have a smaller brain, a shorter life span, and different physiology than humans. To overcome some of these problems, large-animal models, including sheep, mini-pigs, and a number of nonhuman primates are being developed to study the mechanisms of disease and the testing of drugs for human therapies.

Transgenic sheep are one of the most useful large-animal models currently available. One such HD model was constructed using sheep carrying a full-length human *HTT* gene with 73 CAG repeats. Given the relatively long life span of sheep (>10 years), this model makes it possible to study the long-term development of HD, something that cannot be done in mice, where the life span is only 2–3 years. In addition, the size and structure of sheep brains are more similar to those of humans. This makes it possible to do MRI and PET scans that can be directly compared to the results of scans in humans affected with HD.

Continued development of large-animal models will hopefully lead to a more detailed understanding of the mechanism of HD and the development of effective therapies.

ST 6.5 Cellular and Molecular Approaches to Therapy

Because the mutant HTT protein affects a large number of cellular processes through protein–protein interactions and the accumulation of mHTT aggregates, researchers are using multiple approaches to investigate treatment strategies. We will now briefly review several of these approaches.

Stem Cells for Transplantation

Stem cells are undifferentiated somatic cells with two properties: (a) the ability to renew their numbers by mitosis, and (b) the ability to differentiate and form tissue-specific specialized cell types. Research on HD uses human embryonic stem cells (hESCs), derived from the inner cell mass of early embryos, and induced pluripotent stem cells (iPSCs), which can be produced through genetic reprogramming of adult skin cells. Stem cells are used in research on HD for studies of disease mechanisms, drug screening, drug testing, genetic correction of mHTT accumulation, and as donor cells for transplantation.

HD is associated with transcriptional repression of many gene sets, including those essential for survival of nerve cells. One of these repressed genes encodes brain-derived neurotrophic factor (BDNF), which is essential for the survival and function of cells in the striatum. Loss of this factor contributes to the death of striatal brain cells in HD. Human mesenchymal stem cells (MSCs) can be genetically programmed to overexpress BDNF. Injection of human MSCs modified to overexpress BDNF into the brains of a transgenic mouse strain carrying a full-length human *HTT* gene with 128 CAG repeats significantly increased the production of new nerve cells, increased life span, and reduced HD-associated behaviors. Further development of MSCs as a delivery system in other mouse models and large-animal models will be required before the system is ready for human clinical trials, which are expected to begin as soon as additional animal studies are completed.

Identifying Potential Drugs for Therapy

Knowledge of the molecular pathways that trigger cellular degeneration and death in HD offers opportunities to develop drugs that can inhibit or reverse the formation of toxic aggregates of the mHTT protein. One approach is to search for drugs that stimulate refolding of mHTT into a less toxic form by upregulating expression of the protein-folding chaperones present in the cell. A search for activators of HSF1, a transcription factor that regulates chaperone synthesis, screened more than 900,000 compounds. Three candidate compounds were identified, and each was tested in a transgenic rat cell line expressing a mutant human allele with 74 CAG repeats. Each compound caused only a modest upregulation of HSF1, but resulted in a two- to three-fold reduction in levels of mutant protein aggregates. This suggests that even small changes in gene expression can have therapeutic effects.

Instead of reducing levels of the mHTT protein after aggregates have formed, other researchers seek to identify compounds that reduce aggregate formation. A high-throughput binding assay screening identified candidate molecules that inhibited the *in vitro* aggregation of mHTT. Each compound was tested in transgenic cell lines to confirm the findings before beginning studies in transgenic mice. Further pre-clinical testing in cell lines and animal models will be needed before human clinical trials begin.

Although these approaches have been successful in identifying potential therapeutic molecules that can initiate refolding of the mutant protein or prevent aggregation, methods to reduce or eliminate synthesis of mHTT protein would represent an approach that would be a major step forward in developing an effective therapy for HD.

Gene Silencing to Reduce mHTT Levels

Because mHTT initiates a chain of events that lead to cell death and the onset of HD symptoms, therapies that reduce or eliminate the expression or accumulation of mHTT, either alone or in combination with other therapies, might be especially effective in treating HD.

A therapy using synthetic zinc-finger nucleases (ZFNs) to repress transcription of m*HTT* alleles (see Special Topics Chapter 1—CRISPR-Cas and Genome Editing for a discussion of ZFNs) in mouse neuronal cell lines and a cell line from an individual with HD carrying a mutant allele with 45 CAG repeats showed that ZFN treatment significantly repressed m*HTT* expression with no effects on expression of other genes containing CAG repeats.

The ZFN construct was then tested on a transgenic mouse strain carrying 115–160 CAG repeats by injection into the striatal region of one side of the brain. A control vector was injected into the other side of the brain. Two weeks later, levels of mRNA from the normal and mutant alleles were assayed in tissue from each side of the brain. Overall, on the ZFN-injected side, there was a 40–60 percent reduction of m*HTT* mRNA compared with the control side, with no effect on expression of the normal allele on either side (**Figure ST 6.6**). Mice injected on both sides of the brain with the ZFN construct showed no differences in behavioral tests compared with normal mice. This was an important proof of principle that ZFN repression of m*HTT* transcription reduces m*HTT* mRNA and protein levels and may be an effective therapy for HD.

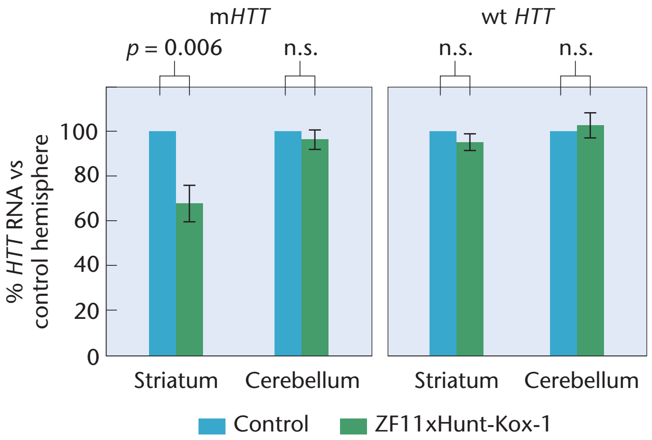


FIGURE ST 6.6 Zinc-finger repression of m*HTT* expression in transgenic mice. Left panel: Expression of m*HTT* is repressed by ~40 percent by injection of the ZFN construct into the striatum, but has no effect on expression in the noninjected cerebellum. The *p* value of 0.006 indicates there is a significant difference between values in the striatum. The abbreviation n.s. means there is no significant difference in the values shown. Right panel: The expression of the normal *HTT* allele is unaffected in both brain regions.

Instead of inhibiting transcription, gene-silencing techniques can be used to intervene in gene expression after transcription but before translation takes place. Two widely used methods of gene silencing use antisense oligonucleotides (ASOs) and RNA interference (RNAi). ASOs are short single-stranded DNAs (8–50 nucleotides long) that bind to target mRNAs by complementary base pairing (**Figure ST 6.7**). The mRNA strand of the resulting DNA–RNA hybrid is degraded by RNase H, a cytoplasmic enzyme. The intact ASO is released and can bind other copies of the target mRNA, marking them for degradation by RNase H.

A ground-breaking study by Lu and Yang transfused an ASO complementary to a human m*HTT* allele into the cerebrospinal fluid of transgenic mice carrying a full-length human m*HTT* allele with 97 CAG repeats. mHTT RNA and mutant *HTT* protein levels were selectively reduced for up to 12 weeks, with a rebound to near normal levels by 16 weeks. With two weeks of continuous infusion at the age of 6 months, treated mice still showed improved motor coordination and behavior 9 months after infusion and 5 months after m*HTT* mRNA and protein levels rebounded to pretreatment levels. Similar ASO studies in nonhuman primates showed a 25–68 percent reduction in levels of *HTT* mRNA in brain regions involved in HD, with no adverse effects.

Human Phase I clinical trials on ASO therapy for HD began in 2015 in Canada and Europe. The drug is being infused into the cerebrospinal fluid, which surrounds and bathes the brain. This trial is designed to evaluate the

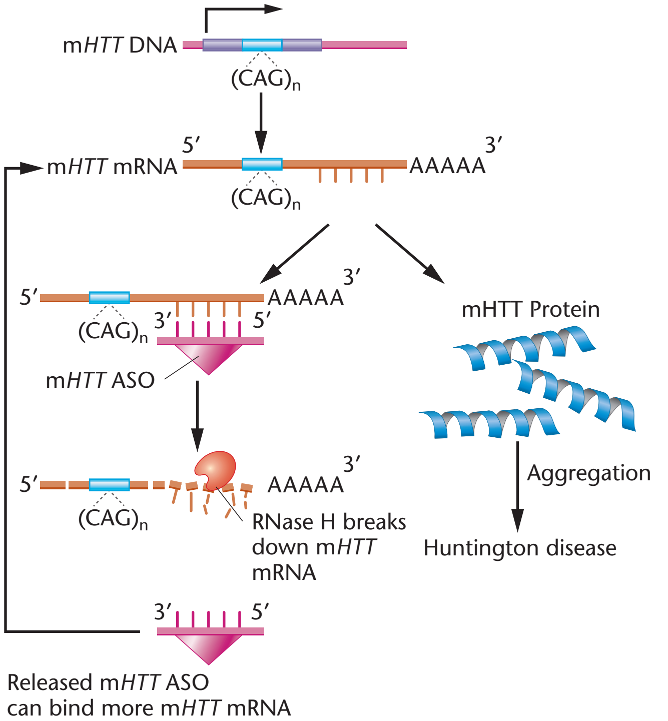


FIGURE ST 6.7 An ASO constructed to bind to m*HTT* mRNA forms a DNA–RNA hybrid, which then attracts RNase to degrade the m*HTT* mRNA, inhibiting translation of the HTT protein and halting disease progression. The released ASO is free to bind other m*HTT* mRNAs and continue the cycle of degradation.

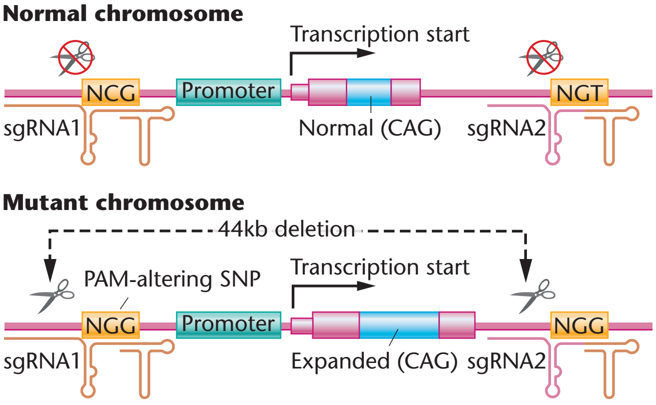


FIGURE ST 6.8 Identification of SNPs that alter the PAM sequences on the chromosome carrying the mutant HD allele permits the design of allele-specific sgRNAs that guide the CRISPR-Cas9 complex to the mutant locus. Cutting by Cas9 excises and silences only the mutant allele. This method can, in theory, be used in a program of patient-specific therapy for HD and other monogenic diseases.

safety of ASO infusion. If this and subsequent clinical trials are successful, the first treatment to directly target the cause of HD will soon be available for the thousands of people affected with this devastating disease.

Gene Editing in Huntington Disease

Over the last decade, there has been rapid development of methods for gene editing. While details of gene editing techniques differ, conceptually they all work in a similar way. A nuclease is guided to cut a specific DNA sequence, which then allows for the replacement or deletion of all or part of a given gene. Because of its specificity and ease of use, the CRISPR-Cas9 system (see Special Topics Chapter 1—CRISPR-Cas and Genome Editing for a discussion of gene editing) is the most widely used method. The enzyme Cas9 can be directed to cut DNA at specific sequences at nearly any location in the genome by means of a single-stranded guide RNA (sgRNA). The guide RNA is complementary to the sequence to be cut and directs cutting by the Cas9 nuclease. Since HD is a dominant disorder, in theory, this disorder can be treated by using this technology to edit and silence the mutant allele.

Using CRISPR/Cas9 technology and human cells carrying a mutant and a normal *HTT* allele, Jong-Min Lee and colleagues edited the disease-associated mutant *HTT* allele, while leaving the normal allele intact. This allele-specific gene editing strategy takes advantage of the fact that Cas9 cleavage depends on the presence of a short nucleotide sequence (5′-NGG-3′) called a protospacer adjacent motif (PAM). To target the mutant allele, the research team identified single nucleotide polymorphisms (SNPs) that modify PAM sites in the mutant allele but not in the normal allele in the HD cell line. With this information in hand, sgRNAs were designed to recognize PAM sites adjacent to the mutant allele (**Figure ST 6.8**). Cutting with Cas9 at these sites removed a 44-kb DNA fragment, completely inactivating the mutant allele, while leaving the normal allele intact. This experiment shows that inactivation of a disease gene can be individually tailored to edit a mutant allele carried by an affected individual and, in theory, can be used to inactivate disease alleles of any gene by editing.

A research team led by Xiao-Jiang Li used CRISPR gene editing to reverse symptoms of HD in transgenic mice carrying exon 1 of a mutant human *HTT* allele containing 140 CAG repeats. They injected one vector carrying Cas9 into the striatum along with another vector that carried sgRNAs to direct Cas9 to target the mutant *HTT* allele to edit out the expanded CAG repeat region. Three weeks after injection, analysis of striatal cells showed that production of the mutant human HTT protein was suppressed and the number of aggregated protein clumps was reduced.

In subsequent experiments, sgRNAs and Cas9 were injected into 9-month-old mice. Over the next 3 months, these mice showed improvements in motor skills including balance, mobility, and muscle coordination. In addition, increases in motor skills were correlated with the amount of aggregated protein cleared from striatal cells.

These results are encouraging, but before CRISPR-Cas9 gene editing can be used in humans, further work in model systems and the elimination of potential safety problems are needed. However, gene editing technology offers the possibility that a cure for HD and many other genetic disorders is not far off.

ST 6.6 The Relationship between HD and Other Neurodegenerative Disorders

Huntington disease is one of 14 known trinucleotide repeat diseases. Nine of these, including HD, are caused by expansion of CAG repeats and the insertion of long polyglutamine (polyQ) tracts in the protein product. Each of these diseases is associated with aggregation of polyQ proteins and the slow but progressive degeneration and death of cells in specific regions of the nervous system. Each of these diseases also begins in midlife and has similar major symptoms.

Trinucleotide diseases show **genetic anticipation**—an earlier age of onset and increasingly severe symptoms in cases where repeat lengths increase from generation to generation (see Chapter 4 for a discussion of anticipation). As with some other polyQ diseases, anticipation in HD is strongly related to paternal inheritance of the mutant allele and is associated with very large numbers of CAG repeats (usually 60 or more). Anticipation resulting from paternal inheritance is also more likely in cases where the father received the mutant allele from his mother. This form of anticipation is thought to be coupled to epigenetically controlled methylation that occurs during imprinting (see Chapter 19 for a discussion of imprinting). However, unlike some other trinucleotide disorders, in HD there is no direct connection between increased repeat length and earlier age of onset, and CAG repeat length accounts for only about 60 percent of the variation in the age of onset associated with anticipation. Evidence suggests that other genes and environmental factors play important roles in determining age of onset when repeat lengths increase.

Most of the research on polyQ diseases has been focused on HD, and therapies for HD are now in pre-clinical and clinical trials. Given the many similarities among polyQ disorders, it seems fair to ask whether therapies for HD might also be effective in treating other polyQ diseases. In a screen for small molecule activators of HSF1, a transcription factor that regulates expression of numerous chaperones, researchers identified a compound they called HSF1A that was a strong activator of HSF1. HSF1A was tested in a transgenic rodent cell line expressing exon 1 of the human *HTT* gene containing 74 CAG repeats. A four-fold reduction in the number of cells containing mutant protein aggregates was observed. Further work showed that the protein chaperones activated by HSF1 acted on the oligomer precursors rather than the larger aggregates. Might this approach be applicable to other disorders? Using a transgenic *Drosophila* model of spinocerebellar ataxia 3 (*SCA 3*) carrying a mutant allele with 78 CAG repeats, researchers showed that successfully reducing aggregation of the mutant SCA3 protein reduces its cytotoxic effects in a transgenic animal model. The fact that this approach works in more than one polyQ-protein disease offers hope that therapies that work in one polyQ disease may be effective in others.

In addition to similarities to other polyQ disorders, HD also shares some features with some non-polyQ neurodegenerative diseases, including Alzheimer disease (AD) and Parkinson disease (PD). Each of these incurable disorders is characterized by the same symptoms seen in Huntington disease, adult onset, the formation of misfolded protein aggregates, neurodegeneration, and disease-specific behavioral changes (see Box 4).

BOX 4 Huntington Disease and Behavior

Huntington disease is an inherited neurodegenerative disorder that destroys specific regions of the brain, resulting in progressive and predictable changes in behavior. This association between behavioral changes and affected brain regions establishes that these regions are directly involved in controlling motor disorders, cognitive decline, and psychiatric problems.

The brain region that suffers the greatest loss of neurons is the caudate nucleus, which is part of the striatum. The caudate nucleus organizes and regulates information flow to the frontal lobes via nerves connecting these two regions. As the caudate region deteriorates, connections to the frontal lobes are destroyed, and the functions of the frontal lobes are compromised. As a result, affected individuals suffer increasingly serious cognitive failures, which ultimately progress to dementia. Eventually, with the decline of frontal lobe functions, apathy and depression are soon evident.

Cell loss in the striatum is progressive, and eventually spreads to other regions of the brain. Over time, brain weight loss can eventually reach 30 percent. Huntington disease progresses gradually, with an average of 10–15 years from the time of diagnosis to death.

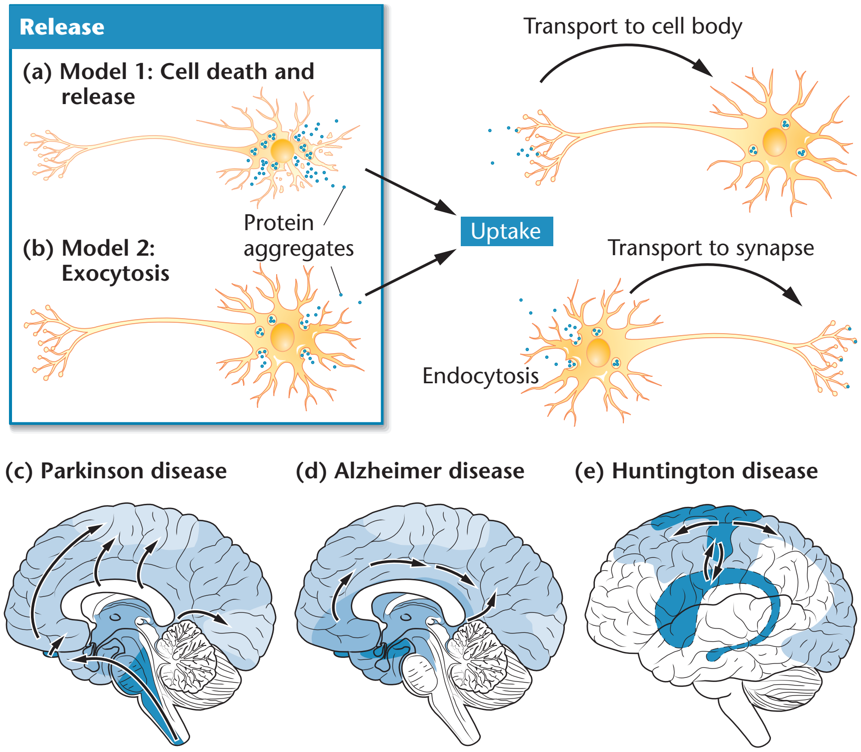


FIGURE ST 6.9 Two potential pathways of release and uptake of protein aggregates by cells of the brain. (a) Passive release as a by-product of cell death can result in uptake by adjacent cells. (b) Experimental evidence shows that exocytosis releases aggregates, which are transferred to nearby uninfected cells by endocytosis. Once inside the cells, vesicle rupture releases the protein aggregates, which then disrupt cellular functions, causing cell death. Darker shading represents brain regions where aggregates first appear. (c) In Parkinson disease, aggregates of alpha-synuclein first appear in the brainstem (darkest shading) and spread from there to occupy most other brain regions. (d) In Alzheimer disease, neurofibrillary tangles and amyloid plaques first appear in the hippocampus and associated regions (darkest shading) and spread to adjacent regions (lighter shadings) that are correlated with progression of symptoms. (e) In Huntington disease, aggregates of mutant huntingtin first appear in the caudate nucleus and nearby basal ganglia (darkest shading), leading to degeneration of these regions. From there, aggregates spread to other brain regions.

However, different proteins form aggregates in each of these disorders: tau in Alzheimer disease, huntingtin in Huntington disease, and alpha-synuclein in Parkinson disease; and aggregates first appear in different brain regions. However, once aggregates form, they spread to other regions of the brain by axonal transport. At least two shared mechanisms may underlie the spread of aggregates and progression in all these diseases, passive release/uptake after cell death or exocytosis/endocytosis (**Figure ST 6.9**).

The alpha-synuclein aggregates associated with Parkinson disease enter unaffected cells by endocytosis, forming intracellular vesicles. The vesicles subsequently rupture, releasing the aggregates into the cytoplasm where they disrupt cellular functions, leading to cell death.

Extensions of this work showed that the protein aggregates in Alzheimer disease and HD also enter unaffected cells by endocytosis and rupture vesicles, accelerating the spread of misfolded protein aggregates and cell death. The finding that all three neurodegenerative diseases use the same mechanism for transcellular spreading of aggregates suggests that an effective therapy for one of these disorders may work for the others.

Looking back, the focus on researching the cause of HD began with a foundation established by a family with a history of HD. Efforts rapidly expanded into a large-scale international program that pioneered the use of genetics and molecular genetics to identify an expanded stretch of polyglutamines in the huntingtin protein as the cause of this disorder. Along the way, researchers developed an integrative strategy that combined old and new methods including pedigree analysis, RFLP markers, somatic cell genetics, Southern blots, new cloning vectors, predictive genetic testing, population genetics, and other methods that are now universally used in genetic research. Nevertheless, in spite of the progress made in more than 30 years of work, there are still no therapies that can halt, reverse, or prevent the onset and progression of this devastating neurogenetic disorder. Looking forward, recent results from the combined use of human cells, animal models, gene editing, and clinical trials suggest that this last barrier may fall in the near future. If this is the case, the approach used in HD research will stand as a paradigm for understanding the structure and function of the brain and the development of therapeutic methods for a range of genetic disorders.

Selected Readings and Resources

Journal Articles

Bates, G. (2005). The molecular genetics of Huntington disease—a history. *Nature Reviews Genetics* 6:766–773.

Botstein, D., White, R.L., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314-331.

Gusella, J., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:234–238.

Huang, W-J., Chen, W-W., and Zhang, X. (2016). Huntington's disease: Molecular basis of pathology and status of current therapeutic approaches (Review). *Exp. Therap. Med.* 12:1951–1956.

Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable in Huntington's disease chromosomes. *Cell* 72:971–983.

Jenkins, J. B., and Conneally, P. M. (1989). The paradigm of Huntington's disease. *Am. J. Hum. Genet.* 45:169–175.

Labbadia, J., and Morimoto, R. I. (2013). Huntington's disease: underlying molecular mechanisms and emerging concepts. *Trends in Biochem. Sci.* 38:378–385.

Lu, X-H. and Yang, X-W. (2012). "Huntingtin holiday": Progress toward an antisense therapy of Huntington's disease. *Neuron* 74(6):964–966.

Shin, J.W., et al. (2016). Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. *Hum. Mol. Genet*. 25:4566–4576.

Wexler, N. (2012). Huntington's disease: Advocacy driving science. *Annu. Rev. Med.* 63:1–22.

Yamamoto, A., Lucas, J.J., and Hen, R. (2000). Reversal of neuropathy and motor dysfunction in a conditional model of Huntington's disease. *Cell* 101(1):57–66.

Yang, S., et al. (2017). CRISPR/Cas9-mediated gene editing ameliorates neurotoxicity in mouse model of Huntington's disease. *J. Clin. Invest*. 127(1):2719–2724.

Web Sites

HOPES Huntington's Disease Outreach Project for Education at Stanford. https://web.stanford.edu/groups/cgi-bin/hopes.test/

Huntington's Disease Society of America. https://www.hdsa.org

International Huntington Association: A Global Effort to Help People with Huntington Disease. https://www.huntington-disease.org

Huntington Study Group. http://www.huntingtonstudygroup.org

Review Questions

1. What are RFLP markers and how were they used to identify which chromosome carries the gene for Huntington disease?
2. Why was information from Nancy Wexler's large pedigrees necessary in locating the *HTT* gene?
3. How do aggregates of mHTT protein form?
4. Why are the results from the inducible mouse model of HD so important?
5. Based on the results from mouse models, is it necessary to have the whole mutant protein (mHTT) present to generate protein aggregates and death of brain cells?
6. What do the results from creating transgenic mice carrying only exon 1 of the mutant allele tell us about the disease?
7. What steps lead from the binding of the mHTT protein to mitochondrial fragmentation and the triggering of apoptosis?
8. Summarize the approaches to therapy designed to reduce the level of the mHTT protein in cells. How can this therapy be delivered to target cells?

Discussion Questions

1. There are nine known progressive neurodegenerative disorders that all share expanded numbers of the CAG codon, which inserts extra glutamine residues into the coding regions of specific genes. Genes carrying such mutations are typically gain-of-function mutations and often share a common mechanism of pathogenesis. Why would such genes be gain-of-function? Speculate on why such diseases may be caused by a common mechanism.
2. Most of the research efforts on Huntington disease have focused on the mechanisms by which the mutant form of the HTT protein causes cell death or the clinical symptoms of this disorder. For the development of effective therapies, how important is it to understand the functions of the normal HTT protein?
3. If we now know that protein aggregates can spread from cell to cell, should the use of stem cell transplants be reconsidered?
4. Why is there an inverse correlation between the number of CAG repeats in a mutant allele and the onset of symptoms?
5. Discuss the ethical issues raised by the use a diagnostic test for those at risk for HD.

This page intentionally left blank

Appendix A

Selected Readings

CHAPTER 1 Introduction to Genetics

Amman, N. H. 2008. In defense of GM crops. *Science* 322: 1465–1466.

Bilen, J., and Bonini, N. M. 2005. *Drosophila* as a model for human neurodegenerative disease. *Annu. Rev. Genet*. 39: 153–171.

Chen, M., Shelton, A., and Ye, G. Y. 2011. Insect-resistant genetically-modified rice in China: From research to commercialization. *Annu. Rev. Entomol*. 56: 81–101.

Cohen, J. C., and Hobbs, H. H. 2013. Simple genetics for a complex disease. *Science* 340: 689–690.

Daya, S., and Berns, K. I. 2008. Gene therapy using adeno-associated virus vectors. *Clin. Microbiol. Rev.* 21: 83–93.

Doudna, J. and Charpentier, E. 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346: 1258096.

Ehrnhoefer, D. E., Butland, S. L., Pouladi, M. A., and Hayden, M. R. 2009. Mouse models of Huntington disease: Variations on a theme. *Dis. Models Mech*. 2: 123–129.

Gurdon, J. B. 2013. The egg and nucleus: A battle for supremacy. *Develop*. 140: 2449–2456.

Lander, E.S. 2016. The heroes of CRISPR. *Cell* 164: 18–28.

Müller, B., and Grossnicklaus, U. 2010. Model organisms—A historical perspective. *J. Proteomics* 73: 2054–2063.

Pearson, H. 2006. What is a gene? *Nature* 441: 399–401.

Reid, J. B., and Ross, J. J. 2011. Mendel's genes: Toward a full molecular characterization. *Genetics* 189: 3–10.

Tiscornia, G., Vivas, E. L., and Belmonte, J. C. 2011. Diseases in a dish: Modeling human genetic disorders using induced pluripotent cells. *Nat. Med*. 17: 1570–1576.

Wisniewski, J-P., Frange, N., Massonneau, A., and Dumas, C. 2002. Between myth and reality: Genetically modified maize, an example of a sizeable scientific controversy. *Biochimie* 84: 1095–1103.

CHAPTER 2 Mitosis and Meiosis

Bornens, M. 2012. The centrosome in cells and organisms. *Science* 335: 422–426.

Brachet, J., and Mirsky, A. E. 1961. *The cell: Meiosis and mitosis,* Vol. 3. Orlando, FL: Academic Press.

DuPraw, E. J. 1970. *DNA and chromosomes.* New York: Holt, Rinehart & Winston.

Glotzer, M. 2005. The molecular requirements for cytokinesis. *Science* 307: 1735–1739.

Glover, D. M., Gonzalez, C., and Raff, J. W. 1993. The centrosome. *Sci. Am.* (June) 268: 62–68.

Golomb, H. M., and Bahr, G. F. 1971. Scanning electron microscopic observations of surface structures of isolated human chromosomes. *Science* 171: 1024–1026.

Hartwell, L. H., and Karstan, M. B. 1994. Cell cycle control and cancer. *Science* 266: 1821–1828.

Hartwell, L. H., and Weinert, T. A. 1989. Checkpoint controls that ensure the order of cell cycle events. *Science* 246: 629–634.

Ishiguro, K., and Watanabe Y. 2007. Chromosome cohesion in mitosis and meiosis. *J. Cell Sci.* 120: 367–369.

Javerzat, J. P. 2010. Directing the centromere guardian. *Science* 327: 150–151.

Mazia, D. 1961. How cells divide. *Sci. Am.* (Jan.) 205: 101–120.

____. 1974. The cell cycle. *Sci. Am.* (Jan.) 235: 54–64.

Mcintosh, J. R., and McDonald, K. L. 1989. The mitotic spindle. *Sci. Am.* (Oct.) 261: 48–56.

Nature Milestones. 2001. *Cell division.* London: Nature Publishing Group.

Watanabe Y. 2005. Shugoshin: Guardian spirit at the centromere. *Curr. Opin. Cell Biol.* 17: 590–595.

Westergaard, M., and von Wettstein, D. 1972. The synaptinemal complex. *Annu. Rev. Genet.* 6: 71–110.

CHAPTER 3 Mendelian Genetics

Bennett, R. L., et al. 1995. Recommendations for standardized human pedigree nomenclature. *Am. J. Hum. Genet.* 56: 745–752.

Carlson, E. A. 1987. *The gene: A critical history,* 2nd ed. Philadelphia: Saunders.

Dunn, L. C. 1965. *A short history of genetics.* New York: McGraw-Hill.

Henig, R. M. 2001. *The monk in the garden: The lost and found genius of Gregor Mendel, the father of genetics.* New York: Houghton-Mifflin.

Kabeche, L., and Compton, D. A. 2013. Cyclin A regulates kinetochore microtubules to promote faithful chromosome segregation. *Nature* 502: 110–114.

Klein, J. 2000. Johann Mendel's field of dreams. *Genetics* 156: 1–6.

Miller, J. A. 1984. Mendel's peas: A matter of genius or of guile? *Sci. News* 125: 108–109.

Olby, R. C. 1985. *Origins of Mendelism,* 2nd ed. London: Constable.

Orel, V. 1996. *Gregor Mendel: The first geneticist.* Oxford: Oxford University Press.

Peters, J., ed. 1959. *Classic papers in genetics.* Englewood Cliffs, NJ: Prentice-Hall.

Schwartz, J. 2008. *In pursuit of the gene: From Darwin to DNA.* Cambridge, MA: Harvard University Press.

Sokal, R. R., and Rohlf, F. J. 1995. *Biometry,* 3rd ed. New York: W. H. Freeman.

Stern, C., and Sherwood, E. 1966. *The origins of genetics: A Mendel source book.* San Francisco: W. H. Freeman.

Stubbe, H. 1972. *History of genetics: From prehistoric times to the rediscovery of Mendel's laws.* Cambridge, MA: MIT Press.

Sturtevant, A. H. 1965. *A history of genetics.* New York: Harper & Row.

Tschermak-Seysenegg, E. 1951. The rediscovery of Mendel's work. *J. Hered.* 42: 163–172.

Welling, F. 1991. Historical study: Johann Gregor Mendel 1822–1884. *Am. J. Med. Genet.* 40: 1–25.

CHAPTER 4 Extensions of Mendelian Genetics

Bartolomei, M. S., and Tilghman, S. M. 1997. Genomic imprinting in mammals. *Annu. Rev. Genet.* 31: 493–525.

Brink, R. A., ed. 1967. *Heritage from Mendel.* Madison: University of Wisconsin Press.

Bultman, S. J., Michaud, E. J., and Woychik, R. P. 1992. Molecular characterization of the mouse *agouti* locus. *Cell* 71: 1195–1204.

Carlson, E. A. 1987. *The gene: A critical history,* 2nd ed. Philadelphia: Saunders.

Choman, A. 1998. The myoclonic epilepsy and ragged-red fiber mutation provides new insights into human mitochondrial function and genetics. *Am. J. Hum. Genet.* 62: 745–751.

Dunn, L. C. 1966. *A short history of genetics.* New York: McGraw-Hill.

Feil, R., and Khosla, S. 1999. Genomic imprinting in mammals: An interplay between chromatin and DNA methylation. *Trends Genet.* 15: 431.

Foster, M. 1965. Mammalian pigment genetics. *Adv. Genet.* 13: 311–339.

Grant, V. 1975. *Genetics of flowering plants.* New York: Columbia University Press.

Harper, P. S., et al. 1992. Anticipation in myotonic dystrophy: New light on an old problem. *Am. J. Hum. Genet.* 51: 10–16.

Mitchell, M. B., and Mitchell, H. K. 1952. A case of maternal inheritance in *Neurospora crassa. Proc. Natl. Acad. Sci. (USA)* 38: 442–449.

Morgan, T. H. 1910. Sex-limited inheritance in *Drosophila. Science* 32: 120–122.

Peters, J. A., ed. 1959. *Classic papers in genetics.* Englewood Cliffs, NJ: Prentice-Hall.

Phillips, P. C. 1998. The language of gene interaction. *Genetics* 149: 1167–1171.

Race, R. R., and Sanger, R. 1975. *Blood groups in man,* 6th ed. Oxford: Blackwell.

Sapienza, C. 1990. Parental imprinting of genes. *Sci. Am.* (Oct.) 363: 52–60.

Siracusa, L. D. 1994. The *agouti* gene: Turned on to yellow. *Cell* 10: 423–428.

Yoshida, A. 1982. Biochemical genetics of the human blood group ABO system. *Am. J. Hum. Genet.* 34: 1–14.

CHAPTER 5 Chromosome Mapping in Eukaryotes

Allen, G. E. 1978. *Thomas Hunt Morgan: The man and his science.* Princeton, NJ: Princeton University Press.

Chaganti, R., Schonberg, S., and German, J. 1974. A manyfold increase in sister chromatid exchange in Bloom syndrome lymphocytes. *Proc. Natl. Acad. Sci.* 71: 4508–4512.

Creighton, H. S., and McClintock, B. 1931. A correlation of cytological and genetical crossing over in *Zea mays. Proc. Natl. Acad. Sci.* 17: 492–497.

Douglas, L., and Novitski, E. 1977. What chance did Mendel's experiments give him of noticing linkage? *Heredity* 38: 253–257.

Ellis, N. A., et al. 1995. The Bloom syndrome gene product is homologous to RecQ helicases. *Cell* 83: 655–666.

Ephrussi, B., and Weiss, M. C. 1969. Hybrid somatic cells. *Sci. Am.* (Apr.) 220: 26–35.

Latt, S. A. 1981. Sister chromatid exchange formation. *Annu. Rev. Genet.* 15: 11–56.

Lindsley, D. L., and Grell, E. H. 1972. *Genetic variations of Drosophila melanogaster*. Washington, DC: Carnegie Institute of Washington.

Morgan, T. H. 1911. An attempt to analyze the constitution of the chromosomes on the basis of sex-linked inheritance in *Drosophila. J. Exp. Zool.* 11: 365–414.

Morton, N. E. 1955. Sequential test for the detection of linkage. *Am. J. Hum. Genet.* 7: 277–318.

____. 1995. LODs—Past and present. *Genetics* 140: 7–12.

Perkins, D. 1962. Crossing over and interference in a multiply marked chromosome arm of *Neurospora. Genetics* 47: 1253–1274.

Ruddle, F. H., and Kucherlapati, R. S. 1974. Hybrid cells and human genes. *Sci. Am.* (July) 231: 36–49.

Stahl, F. W. 1979. *Genetic recombination.* New York: W. H. Freeman.

Stern, C. 1936. Somatic crossing over and segregation in *Drosophila melanogaster. Genetics* 21: 625–631.

Sturtevant, A. H. 1913. The linear arrangement of six sex-linked factors in *Drosophila,* as shown by their mode of association. *J. Exp. Zool.* 14: 43–59.

____. 1965. *A history of genetics.* New York: Harper & Row.

Voeller, B. R., ed. 1968. *The chromosome theory of inheritance: Classical papers in development and heredity.* New York: Appleton-Century-Croft.

Wellcome Trust Case Control Consortium, 2007. Genome association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–676.

Wolff, S., ed. 1982. *Sister chromatid exchange.* New York: Wiley-Interscience.

CHAPTER 6 Genetic Analysis and Mapping in Bacteria and Bacteriophages

Adelberg, E. A. 1960. *Papers on bacterial genetics.* Boston: Little, Brown.

Benzer, S. 1962. The fine structure of the gene. *Sci. Am.* (Jan.) 206: 70–86.

Birge, E. A. 1988. *Bacterial and bacteriophage genetics—An introduction.* New York: Springer-Verlag.

Brock, T. 1990. *The emergence of bacterial genetics.* Cold Spring Harbor, NY: Cold Spring Harbor Press.

Bukhari, A. I., Shapiro, J. A., and Adhya, S. L., eds. 1977. *DNA insertion elements, plasmids, and episomes.* Cold Spring Harbor, NY: Cold Spring Harbor Press.

Cairns, J., Stent, G. S., and Watson, J. D., eds. 1966. *Phage and the origins of molecular biology.* Cold Spring Harbor, NY: Cold Spring Harbor Press.

Campbell, A. M. 1976. How viruses insert their DNA into the DNA of the host cell. *Sci. Am.* (Dec.) 235: 102–113.

Hayes, W. 1968. *The genetics of bacteria and their viruses,* 2nd ed. New York: Wiley.

Hershey, A. D., and Rotman, R. 1949. Genetic recombination between host range and plaque-type mutants of bacteriophage in single cells. *Genetics* 34: 44–71.

Hotchkiss, R. D., and Marmur, J. 1954. Double marker transformations as evidence of linked factors in deoxyribonucleate transforming agents. *Proc. Natl. Acad. Sci. (USA)* 40: 55–60.

Jacob, F., and Wollman, E. L. 1961. Viruses and genes. *Sci. Am.* (June) 204: 92–106.

Kohiyama, M., et al. 2003. Bacterial sex: Playing voyeurs 50 years later. *Science* 301: 802–803.

Kruse, H., and Sorum, H. 1994. Transfer of multiple drug resistance plasmids between bacteria of diverse origins in natural microenvironments. *Appl. Environ. Microbiol.* 60: 4015–4021.

Lederberg, J. 1986. Forty years of genetic recombination in bacteria: A fortieth anniversary reminiscence. *Nature* 324: 627–628.

Luria, S. E., and Delbruck, M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28: 491–511.

Lwoff, A. 1953. Lysogeny. *Bacteriol. Rev.* 17: 269–337.

Miller, J. H. 1992. *A short course in bacterial genetics.* Cold Spring Harbor, NY: Cold Spring Harbor Press.

Miller, R. V. 1998. Bacterial gene swapping in nature. *Sci. Am.* (Jan.) 278: 66–71.

Morse, M. L., Lederberg, E. M., and Lederberg, J. 1956. Transduction in *Escherichia coli* K12. *Genetics* 41: 141–156.

Novick, R. P. 1980. Plasmids. *Sci. Am.* (Dec.) 243: 102–127.

Smith-Keary, P. F. 1989. *Molecular genetics of Escherichia coli.* New York: Guilford Press.

Stahl, F. W. 1987. Genetic recombination. *Sci. Am.* (Nov.) 256: 91–101.

Stent, G. S. 1966. *Papers on bacterial viruses,* 2nd ed. Boston: Little, Brown.

Wollman, E. L., Jacob, F., and Hayes, W. 1956. Conjugation and genetic recombination in *Escherichia coli* K12. *Cold Spring Harb. Symp. Quant. Biol.* 21: 141–162.

Zinder, N. D. 1958. Transduction in bacteria. *Sci. Am.* (Nov.) 199: 38–46.

CHAPTER 7 Sex Determination and Sex Chromosomes

Arbeitman, M. N., et al. 2014. The genetics of sex: exploring differences. *G3: Genomes, Genomics, Genetics* 4: 979–981.

Arnold, A. P., Itoh, Y., and Melamed, E. 2008. A bird's-eye view of sex chromosome dosage compensation. *Annu. Rev. Genomics Hum. Genet.* 9: 109–127.

Court-Brown, W. M. 1968. Males with an XYY sex chromosome complement. *J. Med. Genet.* 5: 341–359.

Davidson, R., Nitowski, H., and Childs, B. 1963. Demonstration of two populations of cells in human females heterozygous for glucose-6-phosphate dehydrogenase variants. *Proc. Natl. Acad. Sci.* (*USA*) 50: 481–485.

Gayen, S., et al. 2016. Sex-specific silencing of X-linked genes by Xist RNA. *Proc. Natl. Acad. Sci.* (*USA*) 113: E309–E318.

Goriely, A., et al. 2012. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human diseases. *Am J. Hum. Genet.* 90: 175–200.

Grath, S., and Parsch, J. 2016. Sex-biased gene expression. *Annu. Rev. Genet.* 50: 29–44.

Hashiyama, K., et al. 2011. *Drosophila* sex lethal gene initiates female development in germline progenitors. *Science* 333: 885–888.

Hodgkin, J. 1990. Sex determination compared in *Drosophila* and *Caenorhabditis. Nature* 344: 721–728.

Holleley, C. E., et al. 2015. Sex reversal triggers the rapid transition from genetic to temperature-dependent sex. *Nature* 523: 79–82.

Hook, E. B. 1973. Behavioral implications of the humans XYY genotype. *Science* 179: 139–150.

Hughes, J. F., et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463: 536–539.

Hughes, J. F., and Rozen, S. 2012. Genomics and genetics of human and primate Y chromosomes. *Annu. Rev. Genom. Hum. Genet.* 13: 83–108.

Irish, E. E. 1996. Regulation of sex determination in maize. *BioEssays* 18: 363–369.

Jacobs, P. A., et al. 1974. A cytogenetic survey of 11,680 newborn infants. *Ann. Hum. Genet.* 37: 359–376.

Kauppi, L., et al. 2011. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science* 331: 916–920.

Koopman, P., et al. 1991. Male development of chromosomally female mice transgenic for Sry. *Nature* 351: 117–121.

Lucchesi, J. 1983. The relationship between gene dosage, gene expression, and sex in *Drosophila. Dev. Genet.* 3: 275–282.

Lyon, M. F. 1972. X-chromosome inactivation and developmental patterns in mammals. *Biol. Rev.* 47: 1–35.

Marshall Graves, J. A. 2009. Birds do it with a Z gene. *Nature* 461: 177–178.

McMillen, M. M. 1979. Differential mortality by sex in fetal and neonatal deaths. *Science* 204: 89–91.

Ng, K., Pullirsch, D., Leeb, M., and Wutz, A. 2007. *Xist* and the order of silencing. *EMBO Reports* 8: 34–39.

Nora, E. P., and Heard, E. 2009. X chromosome inactivation: When dosage counts. *Cell* 139: 865–868.

Penny, G. D., et al. 1996. Requirement for *Xist* in X chromosome inactivation. *Nature* 379: 131–137.

Pieau, C. 1996. Temperature variation and sex determination in reptiles. *BioEssays* 18: 19–26.

Quinn, A. E., et al. 2007. Temperature sex reversal implies sex gene dosage in a reptile. *Science* 316: 411.

Schroeder, A. L. et al. 2016. A novel candidate gene for temperature-dependent sex determination in the common snapping turtle. *Genetics* 203: 557–571.

Smith, C. A., et al. 2009. The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* 461: 267–270.

Stochholm, K., et al. 2012. Criminality in men with Klinefelter's syndrome and XYY syndrome: A cohort study. *BMJ Open* 2: 1–8.

Straub, T., and Becker, P. B. 2007. Dosage compensation: The beginning and end of a generalization. *Nat. Rev. Genet.* 8: 47–57.

Warner, D. A., and Shine, R. 2008. The adaptive significance of temperature-dependent sex determination in a reptile. *Nature* 451: 566–567.

Westergaard, M. 1958. The mechanism of sex determination in dioecious flowering plants. *Adv. Genet.* 9: 217–281.

Witkin, H. A., et al. 1996. Criminality in XYY and XXY men. *Science* 193: 547–555.

Xu, N., Tsai, C-L., and Lee, J. T. 2006. Transient homologous chromosome pairing marks the onset of X inactivation. *Science* 311: 1149–1152.

Yamauchi, Y., et al. 2016. Two genes substitute for the mouse Y chromosome for spermatogenesis and reproduction. *Nature* 351: 514–516.

CHAPTER 8 Chromosome Mutations: Variations in Number and Arrangement

Antonarakis, S. E. 1998. Ten years of genomics, chromosome 21, and Down syndrome. *Genomics* 51: 1–16.

Ashley-Koch, A. E., et al. 1997. Examination of factors associated with instability of the FMR1 CGG repeat. *Am. J. Hum. Genet.* 63: 776–785.

Beasley, J. O. 1942. Meiotic chromosome behavior in species, species hybrids, haploids, and induced polyploids of *Gossypium. Genetics* 27: 25–54.

Blakeslee, A. F. 1934. New jimson weeds from old chromosomes. *J. Hered.* 25: 80–108.

Boue, A. 1985. Cytogenetics of pregnancy wastage. *Adv. Hum. Genet.* 14: 1–58.

Carr, D. H. 1971. Genetic basis of abortion. *Annu. Rev. Genet.* 5: 65–80.

Conrad, D. F., et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704–710.

Croce, C. M. 1996. The FHIT gene at 3p14.2 is abnormal in lung cancer. *Cell* 85: 17–26.

DeArce, M. A., and Kearns, A. 1984. The fragile X syndrome: The patients and their chromosomes. *J. Med. Genet.* 21: 84–91.

Epstein, C. J. 2006. Down's syndrome: Critical genes in a critical region. *Nature* 441: 582–583.

Feldman, M., and Sears, E. R. 1981. The wild gene resources of wheat. *Sci. Am.* (Jan.) 244: 102–112.

Galitski, T., et al. 1999. Ploidy regulation of gene expression. *Science* 285: 251–254.

Gersh, M., et al. 1995. Evidence for a distinct region causing a catlike cry in patients with 5p deletions. *Am. J. Hum. Genet.* 56: 1404–1410.

Hassold, T. J., et al. 1980. Effect of maternal age on autosomal trisomies. *Ann. Hum. Genet.* (*London*) 44: 29–36.

Hassold, T. J., and Hunt, P. 2001. To err (meiotically) is human: The genesis of human aneuploidy. *Nat. Rev. Gen.* 2: 280–291.

Hassold, T., and Jacobs, P. A. 1984. Trisomy in man. *Annu. Rev. Genet.* 18: 69–98.

Hecht, F. 1988. Enigmatic fragile sites on human chromosomes. *Trends Genet.* 4: 121–122.

Hulse, J. H., and Spurgeon, D. 1974. Triticale. *Sci. Am.* (Aug.) 231: 72–81.

Kaiser, P. 1984. Pericentric inversions: Problems and significance for clinical genetics. *Hum. Genet.* 68: 1–47.

Lewis, E. B. 1950. The phenomenon of position effect. *Adv. Genet.* 3: 73–115.

Lewis, W. H., ed. 1980. *Polyploidy: Biological relevance.* New York: Plenum Press.

Lynch, M., and Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1154.

Madan, K. 1995. Paracentric inversions: A review. *Hum. Genet.* 96: 503–515.

Nelson, D. L., and Gibbs, R. A. 2004. The critical region in trisomy 21. *Science* 306: 619–621.

Ohno, S. 1970. *Evolution by gene duplication.* New York: Springer-Verlag.

Oostra, B. A., and Verkerk, A. J. 1992. The fragile X syndrome: Isolation of the FMR-1 gene and characterization of the fragile X mutation. *Chromosoma* 101: 381–387.

Patterson, D. 1987. The causes of Down syndrome. *Sci. Am.* (Aug.) 257: 52–61.

Patterson, D., and Costa, A. 2005. Down syndrome and genetics—A case of linked histories. *Nature Reviews Genetics* 6: 137–145.

Shepard, J., et al. 1983. Genetic transfer in plants through interspecific protoplast fusion. *Science* 21: 683–688.

Shepard, J. F. 1982. The regeneration of potato plants from protoplasts. *Sci. Am.* (May) 246: 154–166.

Stranger, B. E., et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848–854.

Taylor, A. I. 1968. Autosomal trisomy syndromes: A detailed study of 27 cases of Edwards syndrome and 27 cases of Patau syndrome. *J. Med. Genet.* 5: 227–252.

Tjio, J. H., and Levan, A. 1956. The chromosome number of man. *Hereditas* 42: 1–6.

Wilkins, L. E., Brown, J. X., and Wolf, B. 1980. Psychomotor development in 65 home-reared children with cri-du-chat syndrome. *J. Pediatr.* 97: 401–405.

CHAPTER 9 Extranuclear Inheritance

Adams, K. L., et al. 2000. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* 408: 354–357.

Claiborne, A. B., English, R. A., and Kahn, J. P. 2016. Finding an ethical path forward for mitochondrial replacement. *Science* 351:668–670.

Desai, N., et al. 2017. The structure of the yeast mitochondrial ribosome. *Science* 355:528–531.

Freeman, G., and Lundelius, J. W. 1982. The developmental genetics of dextrality and sinistrality in the gastropod *Lymnaea peregra*. *Wilhelm Roux Arch.* 191: 69–83.

Green, B. R., and Burton, H. 1970. Acetabularia chloroplast DNA: Electron microscopic visualization. *Science* 168: 981–982.

He, Y., et al. 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464: 610–614.

Hiendleder, S. 2007. Mitochondrial DNA inheritance after SCNT. *Adv. Exp. Med Biol.* 591: 103–116.

Ishikawa, K. 2008. ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis. *Science* 320: 661–664.

Kang, E. 2016 Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature* 540: 270–275.

Kuroda, R., et al. 2009. Chiral blastomere arrangement dictates zygotic left-right asymmetry pathway in snails. *Science* 462: 790–794.

Lightowlers, R. N., et al. 2015. What is new in mitochondrial disease, and what challenges remain. *Science* 349:1494–1499.

Margulis, L. 1970. *Origin of eukaryotic cells.* New Haven, CT: Yale University Press.

Mitchell, M. B., and Mitchell, H. K. 1952. A case of maternal inheritance in *Neurospora crassa. Proc. Natl. Acad. Sci.* (*USA*) 38: 442–449.

Nüsslein-Volhard, C. 1996. Gradients that organize embryo development. *Sci. Am.* (Aug.) 275: 54–61.

Preer, J. R. 1971. Extrachromosomal inheritance: Hereditary symbionts, mitochondria, chloroplasts. *Annu. Rev. Genet.* 5: 361–406.

Sager, R. 1965. Genes outside the chromosomes. *Sci. Am.* (Jan.) 212: 70–79.

____. 1985. Chloroplast genetics. *BioEssays* 3: 180–184.

Schwartz, R. M., and Dayhoff, M. O. 1978. Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. *Science* 199: 395–403.

Sonneborn, T. M. 1959. Kappa and related particles in *Paramecium. Adv. Virus Res.* 6: 229–256.

Sturtevant, A. H. 1923. Inheritance of the direction of coiling in *Limnaea. Science* 58: 269–270.

Tachibana, M., et al. 2009. Mitochondrial gene replacement in primate offspring and embryonic stem cells. *Nature* 461: 367–372.

Taylor, R. W., and Turnbull, D. M., 2005. Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* 6: 389–402.

Vafai, S. B., and Mootha, V. K. 2012. Mitochondrial disorders as windows into an ancient organelle. *Nature* 491: 374–383.

CHAPTER 10 DNA Structure and Analysis

Adleman, L. M. 1998. Computing with DNA. *Sci. Am.* (Aug.) 279: 54–61.

Alloway, J. L. 1933. Further observations on the use of pneumococcus extracts in effecting transformation of type *in vitro*. *J. Exp. Med.* 57: 265–278.

Avery, O. T., MacLeod, C. M., and McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med*. 79: 137–158. (Reprinted in Taylor, J. H. 1965. *Selected papers in molecular genetics.* Orlando, FL: Academic Press.)

Britten, R. J., and Kohne, D. E. 1968. Repeated sequences in DNA. *Science* 161: 529–540.

Chargaff, E. 1950. Chemical specificity of nucleic acids and mechanism for their enzymatic degradation. *Experientia* 6: 201–209.

Darnell, J. E. 1985. RNA. *Sci. Am.* (Oct.) 253: 68–87.

Dawson, M. H. 1930. The transformation of pneumococcal types: I. The interconvertibility of type-specific *S. pneumococci. J. Exp. Med.* 51: 123–147.

Dickerson, R. E., et al. 1982. The anatomy of A-, B-, and Z-DNA. *Science* 216: 475–485.

Dubos, R. J. 1976. *The professor, the institute and DNA: Oswald T. Avery, his life and scientific achievements.* New York: Rockefeller University Press.

Felsenfeld, G. 1985. DNA. *Sci. Am.* (Oct.) 253: 58–78.

Franklin, R. E., and Gosling, R. G. 1953. Molecular configuration in sodium thymonucleate. *Nature* 171: 740–741.

Griffith, F. 1928. The significance of pneumococcal types. *J. Hyg.* 27: 113–159.

Guthrie, G. D., and Sinsheimer, R. L. 1960. Infection of protoplasts of *Escherichia coli* by subviral particles. *J. Mol. Biol.* 2: 297–305.

Hershey, A. D., and Chase, M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Phys.* 36: 39–56. (Reprinted in Taylor, J. H. 1965. *Selected papers in molecular genetics.* Orlando, FL: Academic Press.)

Levene, P. A., and Simms, H. S. 1926. Nucleic acid structure as determined by electrometric titration data. *J. Biol. Chem.* 70: 327–341.

McCarty, M. 1985. *The transforming principle: Discovering that genes are made of DNA.* New York: W. W. Norton.

Olby, R. 1974. *The path to the double helix*. Seattle: University of Washington Press.

Pauling, L., and Corey, R. B. 1953. A proposed structure for the nucleic acids. *Proc. Natl. Acad. Sci.* (*USA*) 39: 84–97.

Rich, A., Nordheim, A., and Wang, A. H.-J. 1984. The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem*. 53: 791–846.

Spizizen, J. 1957. Infection of protoplasts by disrupted T2 viruses. *Proc. Natl. Acad. Sci.* (*USA*) 43: 694–701.

Varmus, H. 1988. Retroviruses. *Science* 240: 1427–1435.

Watson, J. D. 1968. *The double helix*. New York: Atheneum.

Watson, J. D., and Crick, F. C. 1953a. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acids. *Nature* 171: 737–738.

____. 1953b. Genetic implications of the structure of deoxyribose nucleic acid. *Nature* 171: 964.

Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R. 1953. Molecular structure of desoxypentose nucleic acids. *Nature* 171: 738–740.

CHAPTER 11 DNA Replication and Recombination

Blackburn, E. H. 1991. Structure and function of telomeres. *Nature* 350: 569–572.

DeLucia, P., and Cairns, J. 1969. Isolation of an *E. coli* strain with a mutation affecting DNA polymerase. *Nature* 224: 1164–1166.

Gilbert, D. M. 2001. Making sense of eukaryotic DNA replication origins. *Science* 294: 96–100.

Holliday, R. 1964. A mechanism for gene conversion in fungi. *Genet. Res*. 5: 282–304.

Holmes, F. L. 2001. *Meselson, Stahl, and replication of DNA: A history of the "most beautiful experiment in biology."* New Haven, CT: Yale University Press.

Kornberg, A. 1960. Biological synthesis of DNA. *Science* 131: 1503–1508.

Kornberg, A., and Baker, T. 1992. *DNA replication,* 2nd ed. New York: W. H. Freeman.

Krejci, L., et al. 2012. Homologous recombination and its regulation. *Nucleic Acids Res.* 40: 5795–5818.

Krogh, B. O., and Symington, L. S. 2004. Recombination proteins in yeast. *Annu. Rev. Genet*. 38: 233–271.

Leman, A. R., and Noguchi, E. 2013. The replication fork: Understanding the eukaryotic replication machinery and the challenges to genome duplication. *Genes* (Basel) 4: 1–32.

Luke, B., and Linguer, J. 2009. TERRA: Telomere repeat-containing RNA. *EMBO J.* 28: 2503–2510.

Meselson, M., and Stahl, F. W. 1958. The replication of DNA in *Escherichia coli. Proc. Natl. Acad. Sci.* (*USA*) 44: 671–682.

Okazaki, T., et al. 1979. Structure and metabolism of the RNA primer in the discontinuous replication of prokaryotic DNA. *Cold Spring Harbor Symp. Quant. Biol.* 43: 203–222.

Radman, M., and Wagner, R. 1988. The high fidelity of DNA duplication. *Sci. Am.* (Aug.) 259: 40–46.

____. 1987. Genetic recombination. *Sci. Am.* (Feb.) 256: 90–101.

Redon, S., et al. 2010. The noncoding RNA TERRA is a natural ligand of human telomerase. *Nucleic Acid Res.* 38: 5797–5806.

Shay, J. W., and Wright, W. E. 2011. Role of telomeres and telomerase in cancer. *Semin. Cancer Biol.* 21: 349–353.

Takeda, D. Y., and Dutta, A. 2005. DNA replication and progression through S phase. *Oncogene* 24: 2827–2843.

Taylor, J. H., Woods, P. S., and Hughes, W. C. 1957. The organization and duplication of chromosomes revealed by autoradiographic studies using tritium-labeled thymidine. *Proc. Natl. Acad. Sci.* (*USA*) 48: 122–128.

Wang, J. C. 1987. Recent studies of DNA topoisomerases. *Biochim. Biophys. Acta* 909: 1–9.

Whitehouse, H. L. K. 1982. *Genetic recombination: Understanding the mechanisms*. New York: Wiley.

CHAPTER 12 DNA Organization in Chromosomes

Angelier, N., et al. 1984. Scanning electron microscopy of amphibian lampbrush chromosomes. *Chromosoma* 89: 243–253.

Beerman, W., and Clever, U. 1964. Chromosome puffs. *Sci. Am.* (Apr.) 210: 50–58.

Carbon, J. 1984. Yeast centromeres: Structure and function. *Cell* 37: 352–353.

Chen, T. R., and Ruddle, F. H. 1971. Karyotype analysis utilizing differential stained constitutive heterochromatin of human and murine chromosomes. *Chromosoma* 34: 51–72.

DuPraw, E. J. 1970. *DNA and chromosomes.* New York: Holt, Rinehart & Winston.

Gall, J. G. 1981. Chromosome structure and the C-value paradox. *J. Cell Biol.* 91: 3s–14s.

Hewish, D. R., and Burgoyne, L. 1973. Chromatin substructure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem. Biophys. Res. Comm.* 52: 504–510.

Korenberg, J. R., and Rykowski, M. C. 1988. Human genome organization: Alu, LINES, and the molecular organization of metaphase chromosome bands. *Cell* 53: 391–400.

Kornberg, R. D. 1975. Chromatin structure: A repeating unit of histones and DNA. *Science* 184: 868–871.

Kornberg, R. D., and Klug, A. 1981. The nucleosome. *Sci. Am.* (Feb.) 244: 52–64.

Krogh, B. O., and Symington, L. S. 2004. Recombination proteins in yeast. *Annu. Rev. Genet.* 38: 233–271.

Lorch, Y., et al. 2006. Chromatin remodeling by nucleosome disassembly in vitro. *Proc. Nat. Acad. Sci.* 103: 3090–3093.

Luger, K., et al. 1997. Crystal structure of the nucleosome core particle at 2.8 resolution. *Nature* 389: 251–256.

Moyzis, R. K. 1991. The human telomere. *Sci. Am.* (Aug.) 265: 48–55.

Olins, A. L., and Olins, D. E. 1974. Spheroid chromatin units (v bodies). *Science* 183: 330–332.

____. 1978. Nucleosomes: The structural quantum in chromosomes. *Am. Sci.* 66: 704–711.

Singer, M. F. 1982. SINES and LINES: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28: 433–434.

Takeda, D. Y., and Dutta, A. 2005. DNA replication and progression through S phase. *Oncogene* 24: 2827–2843.

CHAPTER 13 The Genetic Code and Transcription

Barrell, B. G., Air, G., and Hutchinson, C. 1976. Overlapping genes in bacteriophage phi X174 *Nature* 264: 34–40.

Barrell, B. G., Banker, A. T., and Drouin, J. 1979. A different genetic code in human mitochondria. *Nature* 282: 189–194.

Bass, B. L., ed. 2000. *RNA editing*. Oxford: Oxford University Press.

Brenner, S., Jacob, F., and Meselson, M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190: 575–580.

Brenner, S., Stretton, A. O. W., and Kaplan, D. 1965. Genetic code: The nonsense triplets for chain termination and their suppression. *Nature* 206: 994–998.

Cattaneo, R. 1991. Different types of messenger RNA editing. *Annu. Rev. Genet.* 25: 71–88.

Cech, T. R. 1986. RNA as an enzyme. *Sci. Am.* (Nov.) 255(5): 64–75.

____. 1987. The chemistry of self-splicing RNA and RNA enzymes. *Science* 236: 1532–1539.

Chambon, P. 1981. Split genes. *Sci. Am.* (May) 244: 60–71.

Cramer, P., et al. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288: 640–649.

Crick, F. H. C. 1962. The genetic code. *Sci. Am.* (Oct.) 207: 66–77.

____. 1966a. The genetic code: III. *Sci. Am.* (Oct.) 215: 55–63.

____. 1966b. Codon–anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* 19: 548–555.

Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. 1961. General nature of the genetic code for proteins. *Nature* 192: 1227–1232.

Darnell, J. E. 1983. The processing of RNA. *Sci. Am.* (Oct.) 249: 90–100.

Dickerson, R. E. 1983. The DNA helix and how it is read. *Sci. Am.* (Dec.) 249: 94–111.

Dugaiczk, A., et al. 1978. The natural ovalbumin gene contains seven intervening sequences. *Nature* 274: 328–333.

Fica, S. M., et al. 2013. RNA catalyses nuclear pre-mRNA splicing. *Nature* 503: 229–234.

Fiers, W., et al. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature* 260: 500–507.

Gamow, G. 1954. Possible relation between DNA and protein structures. *Nature* 173: 318.

Hall, B. D., and Spiegelman, S. 1961. Sequence complementarity of T2-DNA and T2-specific RNA. *Proc. Natl. Acad. Sci.* (*USA*) 47: 137–146.

Hamkalo, B. 1985. Visualizing transcription in chromosomes. *Trends Genet*. 1: 255–260.

Khorana, H. G. 1967. Polynucleotide synthesis and the genetic code. *Harvey Lectures* 62: 79–105.

Makalowska, I., et al. 2005. Overlapping genes in vertebrate genomes. *Comput. Biol. Chem*. 29: 1–12.

Miller, O. L., Hamkalo, B., and Thomas, C. 1970. Visualization of bacterial genes in action. *Science* 169: 392–395.

Nirenberg, M. W. 1963. The genetic code: II. *Sci. Am.* (Mar.) 190: 80–94.

O'Malley, B., et al. 1979. A comparison of the sequence organization of the chicken ovalbumin and ovomucoid genes. In R. Axel et al., eds., *Eucaryotic gene regulation*, Orlando, FL: Academic Press, pp. 281–299.

Papasaikas, P., and Valcarecel, J. 2016. The spliceosome: The ultimate RNA chaperone and sculptor. *Trends Biochem.* Sci. 41: 33–45.

Ray-Soni,A. Bellecourt, M. J., and Landick, R. 2016. Mechanisms of bacterial transcription termination: All good things must end. *Annu. Rev. Biochem*. 85: 319–347.

Reed, R., and Maniatis, T. 1985. Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell* 41: 95–105.

Rhee, H. S., and Pugh, B. F. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* 483: 205–301.

Ridley, M., 2006. *Francis Crick: Discoverer of the genetic code*. New York: HarperCollins.

Roberts, J. W. 2006. RNA polymerase: A scrunching machine. *Science* 314: 1097–1098.

Sharp, P. A. 1994. Nobel lecture: Split genes and RNA splicing. *Cell* 77: 805–815.

Steitz, J. A. 1988. Snurps. *Sci. Am.* (June) 258(6): 56–63.

Watson, J. D. 1963. Involvement of RNA in the synthesis of proteins. *Science* 140: 17–26.

Woychik, N. A., and Jampsey, M. 2002. The RNA polymerase II machinery: Structure illuminates function. *Cell* 108: 453–464.

CHAPTER 14 Translation and Proteins

Agris, P. F., Narendran, A., Sarachan, K., Väre, V. Y. P., and Eruysal, E. 2017. The importance of being modified: The role of RNA modifications in translational fidelity. *Enzymes* 41: 1–50.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science* 181: 223–230.

Atkins, J. F., and Baranov, P. V. 2007. Translation: Duality in the genetic code. *Nature* 448: 1004–1005.

Beadle, G. W., and Tatum, E. L. 1941. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci.* (*USA*) 27: 499–506.

Beet, E. A. 1949. The genetics of the sickle-cell trait in a Bantu tribe. *Ann. Eugenics* 14: 279–284.

Ben-Sham, A. 2010. Crystal structure of the eukaryotic ribosome. *Science* 330: 1203–1208.

Blau, N. 2016. Genetics of phenylketonuria: Then and now. *Hum. Mutat.* 37: 508–515.

Brenner, S. 1955. Tryptophan biosynthesis in *Salmonella typhimurium*. *Proc. Natl. Acad. Sci.* (*USA*) 41: 862–863.

Doolittle, R. F. 1985. Proteins. *Sci. Am.* (Oct.) 253: 88–99.

Fischer, N., et al. 2010. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature* 466: 329–336.

Frank, J. 1998. How the ribosome works. *Am. Sci.* 86: 428–439.

Garrod, A. E. 1902. The incidence of alkaptonuria: A study in chemical individuality. *Lancet* 2: 1616–1620.

____. 1909. Inborn errors of metabolism. London: Oxford University Press. (Reprinted 1963, Oxford University Press, London.)

Garrod, S. C. 1989. Family influences on A. E. Garrod's thinking. *J. Inher. Metab. Dis.* 12: 2–8.

Hartl, F. U. 2017. Protein misfolding diseases. *Annu. Rev. Biochem*. 86: 21–26.

Ingram, V. M. 1957. Gene mutations in human hemoglobin: The chemical difference between normal and sickle-cell hemoglobin. *Nature* 180: 326–328.

Kaczanowska, M., and Rydén-Aulin, M. 2007. Ribosome biogenesis and the translation process in *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* 71: 477–494.

Khatter, H., Myasnikov, A. G., Natchiar, S. K., and Klaholz., B. P. 2015. Structure of the human 80S ribosome. *Nature* 520: 640–645.

Korostelev, A., and Noller, H. F. 2007. The ribosome in focus: New structures bring new insights. *Trends Biochem. Sci.* 32: 434–441.

Koshland, D. E. 1973. Protein shape and control. *Sci. Am.* (Oct.) 229: 52–64.

Lake, J. A. 1981. The ribosome. *Sci. Am.* (Aug.) 245: 84–97.

Lomakin, I. B., and Steitz, T. A. 2013. The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* 500: 307–311.

Neel, J. V. 1949. The inheritance of sickle-cell anemia. *Science* 110: 64–66.

Nirenberg, M. W., and Leder, P. 1964. RNA codewords and protein synthesis. *Science* 145: 1399–1407.

Nomura, M. 1984. The control of ribosome synthesis. *Sci. Am.* (Jan.) 250: 102–114.

Pauling, L., Itano, H. A., Singer, S. J., and Wells, I. C. 1949. Sickle-cell anemia: A molecular disease. *Science* 110: 543–548.

Ramakrishnan, V. 2002. Ribosome structure and the mechanism of translation. *Cell* 108: 557–572.

Rich, A., and Houkim, S. 1978. The three-dimensional structure of transfer RNA. *Sci. Am.* (Jan.) 238: 52–62.

Rich, A., Warner, J. R., and Goodman, H. M. 1963. The structure and function of polyribosomes. *Cold Spring Harbor Symp. Quant. Biol.* 28: 269–285.

Richards, F. M. 1991. The protein-folding problem. *Sci. Am.* (Jan.) 264: 54–63.

Rould, M. A., et al. 1989. Structure of *E. coli* glutaminyl-tRNA synthetase complexed with tRNAGln and ATP at 2.8 resolution. *Science* 246: 1135–1142.

Srb, A. M., and Horowitz, N. H. 1944. The ornithine cycle in *Neurospora* and its genetic control. *J. Biol. Chem.* 154: 129–139.

Warner, J., and Rich, A. 1964. The number of soluble RNA molecules on reticulocyte polyribosomes. *Proc. Natl. Acad. Sci.* (*USA*) 51: 1134–1141.

Wimberly, B. T., et al. 2000. Structure of the 30S ribosomal subunit. *Nature* 407: 327–333.

Woolford, J. L. Jr., and Baserga, S. J. 2013. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics* 195: 643–681.

Yuan, J., et al. 2010. Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett.* 584: 342–349.

Yusupov, M. M., et al. 2001. Crystal structure of the ribosome at 5.5A resolution. *Science* 292: 883–896.

CHAPTER 15 Gene Mutation, DNA Repair, and Transposition

Arnheim, N., and Calabrese, P. 2009. Understanding what determines the frequency and pattern of human germline mutations. *Nature Rev. Genetics* 10: 478–488.

Cairns, J., Overbaugh, J., and Miller, S. 1988. The origin of mutants. *Nature* 335: 142–145.

Cleaver, J. E. 2005. Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nat. Rev. Cancer* 5: 564–571.

Comfort, N. C. 2001. *The tangled field: Barbara McClintock's search for the patterns of genetic control.* Cambridge, MA: Harvard University Press.

Hancks, D. C., and Kazazian, H. H. 2016. Roles for retrotransposon insertions in human disease. *Mob. DNA* (May 6) 7: 9. doi: 10.1186/s13100-016-0065-9.

Luria, S. E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28(6): 491–511.

Mirkin, S. M. 2007. Expandable DNA repeats and human disease. *Nature* 447: 932–940.

Mortelmans, K., and Zeigler, E. 2000. The Ames Salmonella/microsome mutagenicity assay. *Mut. Res.* 455: 29–60.

O'Driscoll, M., and Jeggo, P. A. 2006. The role of double-strand break repair—insights from human genetics. *Nat. Rev. Genet.*7: 45–51.

O'Hare, K. 1985. The mechanism and control of P element transposition in *Drosophila. Trends Genet.* 1: 250–254.

Ortega-Recalde, O., et al. 2013. Whole-exome sequencing enables rapid determination of xeroderma pigmentosum molecular etiology. *PLOS One* 8: e64692.

Sotero-Caio, C. G., et al. 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol Evol.* 9: 161–177.

CHAPTER 16 Regulation of Gene Expression in Bacteria

Antson, A. A., et al. 1999. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* 401: 235–242.

Beckwith, J. R., and Zipser, D., eds. 1970. *The lactose operon*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Bertrand, K., et al. 1975. New features of the regulation of the tryptophan operon. *Science* 189: 22–26.

Breaker, Ronald R. 2008. Complex riboswitches. *Science* 319: 1795–1797.

Gilbert, W., and Müller-Hill, B. 1966. Isolation of the lac repressor. *Proc. Natl. Acad. Sci.* (*USA*) 56: 1891–1898.

____. 1967. The lac operator is DNA. *Proc. Natl. Acad. Sci.* (*USA*) 58: 2415–2421.

Jacob, F., and Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318–356.

Lewis, M., et al. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 271: 1247–1254.

Maumita, M., et al. 2003. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* 113: 577–586.

Serganov, A. 2009. The long and the short of riboswitches. *Curr. Opin. Struct. Biol.* 19: 251–259.

Stroynowski, I., and Yanofsky, C. 1982. Transcript secondary structures regulate transcription termination at the attenuator of *S. marcescens* tryptophan operon. *Nature* 298: 34–38.

Valbuzzi, A., and Yanofsky, C. 2001. Inhibition of the *B. subtilis* regulatory protein TRAP by the TRAP-inhibitory protein, AT. *Science* 293: 2057–2061.

Yanofsky, C. 1981. Attenuation in the control of expression of bacterial operons. *Nature* 289: 751–758.

CHAPTER 17 Transcriptional Regulation in Eukaryotes

Becker, P. B., and Hörz, W. 2002. ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.* 71: 247–273.

Chen, H., et al. 2015. Functional organization of the human 4D nucleome. *Proc. Natl. Acad. Sci.* (*USA*) 112: 8002–8007.

Ecker, J. R., et al. 2012. ENCODE explained. *Nature* 489: 52–55.

Egriboz, O., et al. 2013. Self-association of the Ga14 inhibitor protein Gal80 is impaired by Gal3: Evidence for a new mechanism in the GAL gene switch. *Mol. Cell Biol.* 33: 3667–3674.

Feuerborn, A., and Cook, P. R. 2015. Why the activity of a gene depends on its neighbors. *Trends Genet.* 31: 483–490.

Juven-Gershon, T., and Kadonaga, J. T. 2009. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* 339: 225–229.

Lee, T. I., and Young, R. A. 2013. Transcriptional regulation and its misregulation in disease. *Cell* 152: 1237–1251.

Lorch, Y., Maier-Davis, B., and Kornberg, R. D. 2010. Mechanism of chromatin remodeling. *Proc. Natl. Acad. Sci.* (*USA*) 107: 3458–3462.

Pennacchio, L. A., et al. 2013. Enhancers: Five essential questions. *Nature Rev. Genet.* 14: 288–295.

Rieder, D., Trajanoski, Z., and McNally, J. G. 2012. Transcription factories. *Front. Genet.* 3: 221.

Weintraub, H., and Groudine, M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* 193: 848–856.

CHAPTER 18 Posttranscriptional Regulation in Eukaryotes

Barrett, S. P., and Salzman, J. 2016. Circular RNAs: Analysis, expression and potential functions. *Development* 143: 1838–1847.

Bobbin, M. L., and Rossi, J. J. 2016. RNA Interference (RNAi)-based therapeutics: Delivering on the promise? *Annu. Rev. Pharmacol. Toxicol.* 56: 103–122.

Cesana, M., et al. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147: 358–369.

Eliscovich, C., Buxbaum, A. R., Katz, Z. B., and Singer, R. H. 2013. mRNA on the move: The road to its biological destiny. *J. Biol. Chem.* 288: 20361–20368.

Farrar, M. A., et al. 2017. Emerging therapies and challenges in spinal muscular atrophy. *Ann. Neurol.* 81: 355–368.

Fire, A. et al. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806–811.

Fu, X., and Ares Jr, M. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Rev. Genetics* 15: 689–701.

Iwakawa, H., and Tomari, Y. 2015. The functions of microRNAs: mRNA decay and translational repression. *Trends Cell Biol.* 25: 651–665.

Lee, Y., and Rio, D. C. 2015. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* 84: 291–323.

Licatalosi, D. D., and Darnell, R. B. 2010. RNA processing and its regulation: global insights into biological networks. *Nature Rev. Genetics* 11: 75–87.

Newbury, S. F. 2006. Control of mRNA stability in eukaryotes. *Biochem. Soc. Trans.* 34: 30–34.

Rana, T. M. 2007. Illuminating the silence: Understanding the structure and function of small RNAs. *Nature Mol. Cell Biol.* 8: 23–34.

Scotti, M. M., and Swanson, M. S. 2016. RNA mis-splicing in disease. *Nature Rev. Genetics* 17: 19–32.

Wheeler, T. M., and Thornton, C. A. 2007. Myotonic dystrophy: RNA-mediated muscle disease. *Curr. Opin. Neurol.* 20: 572–576.

Zheng, N., and Shabek, N. 2017. Ubiquitin ligases: Structure, function, and regulation. *Annu. Rev. Biochem.* 86: 129–157.

CHAPTER 19 Epigenetics

Chess, A. 2013. Random and non-random monoallelic expression. *Neuropsychopharm. Rev*. 38: 55–61.

Cubas, P., Vincent, C., and Coen, E. 1999. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* 401: 157–161.

Deans, C., and Maggert, D. 2015. What do you mean, "epigenetic"? *Genetics* 199: 887–896.

Eckersley-Maslin, M. A., and Spector, D. L. 2014. Random monoallelic expression: Regulating gene expression one allele at a time. *Trends Genet*. 30: 237–244.

Kalish, J. M., Jiang, C., and Bartolomei, M. S. 2014. Epigenetics and imprinting in human disease. *Int. J. Dev. Biol.* 58: 291–298.

Lee, J-S., Smith, E., and Shilatifard, A. 2010. The language of histone crosstalk. *Cell* 142: 682–685.

Lokk, K., et al. 2014. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*. 15; r54.

Lorch, Y., et al., 2010. Mechanism of chromatin remodeling. *Proc. Nat. Acad. Sci*. 107: 3458–3462.

Messerschmidt, D. M., Knowles, B., and Solter, D. 2014. DNA methylation dynamics during epigenetic reprogramming in the germ line and preimplantation embryos. *Genes Dev*. 28: 812–828.

Miska, E. A., and Ferguson-Smith, A. C. 2016. Transgenerational inheritance: Models and mechanisms of non-DNA sequence-based inheritance. *Science* 354: 59–63.

Odum L. N., and Segars, J. 2010. Imprinting disorders and assisted reproductive technologies. *Curr. Opin. Endocrinol. Diabetes Obes.* 17: 517–522.

Reinius B., and Sandberg, R. 2015. Random monoallelic expression of autosomal genes: Stochastic transcription and allele-level regulation. *Nat. Rev. Genet*. 16: 653–662.

Rinn, J. L., and Chang, H. Y. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81: 145–166.

Rivera, C. M., and Ren, B. 2013. Mapping human epigenomes. *Cell* 155: 39–55.

Skinner, M. K. 2014. A new kind of inheritance. *Sci. Am*. 311 (Aug); 44–51.

Szyf, M. 2008. The role of DNA hypermethylation and demethylation in cancer and cancer therapy. *Curr. Oncol*. 15: 72–75.

Veenendaal, M. V. E., et al. 2013. Transgenerational effects of prenatal exposure to the 1944–45 Dutch famine. *BJOG* 120: 548–554.

Youngson, N. A., and Whitelaw, E. 2008. Transgenerational epigenetic effects. *Annu. Rev. Genomics Hum. Genet*. 9: 233–237.

CHAPTER 20 Recombinant DNA Technology

Barrangou, R., and Doudna, J. A. 2016. Applications of CRISPR technologies in research and beyond. *Nature Biotech*. 34:933–941.

Brownlee, C. 2005. Danna and Nathans: Restriction enzymes and the boon to modern molecular biology. *Proc. Nat. Acad. Sci.* 103: 5909.

Levy, S. E., and Myers, R. M. 2016. Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* 17: 95–115.

____. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470: 198–203.

Mullis, K. B. 1990. The unusual origin of the polymerase chain reaction. *Sci. Am.* (Apr.) 262: 56–65.

Rothberg, J. M., and Leamon, J. H. 2008. The development and impact of 454 sequencing. *Nat. Biotechnol.* 26: 1117–1124.

Saleheen, D. et al 2017. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 544: 235–239.

Sanger, F., et al. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* (*USA*) 74: 5463–5467.

Shendure, J., and Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnol.* 26: 1135–1145.

Shendure, J., et al. 2017. DNA sequencing at 40: past, present and future. *Nature*, 550: 345–353.

Snyder, M., Du, J., and Gerstein, M. 2010. Personal genome sequencing: Current approaches and challenges. *Genes Dev.* 24: 423–431.

Southern, E. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98: 503–507.

CHAPTER 21 Genomic Analysis

Abersold, R., and Mann, M. 2016. Mass-spectromic exploration of proteome structure and function. *Nature* 537: 347–355.

Asara, J. M., et al. 2007. Protein sequences from mastodon and *Tyrannosaurus Rex* revealed by mass spectrometry. *Science* 316: 280–285.

Campbell, K. L., and Hofreiter, M. 2012. New life for ancient DNA. *Sci. Am.* 307: 46–51.

Conrad, D. F., et al. 2010. Origins and functional impact of copy number variations in the human genome. *Nature* 464: 704–712.

Doolittle, W. F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci.* (*USA*) 110: 5294–5300.

Green, R. E., et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134: 416–426.

Green, R. E., et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722.

Grice, E. A., and Segre, J. A. 2012. The human microbiome: Our second genome. *Annu. Rev. Genomics Hum. Genetic.* 13: 151–170.

Hawkins, R. D., Hon, G. C., and Ren, B. 2010. Next-generation genomics: An integrative approach. *Nature Rev. Genet.* 11: 476–486.

Hoskins, R. A., et al. 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 316: 1625–1628.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

____. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945.

Joyce, A. R., and Palsson, B. O. 2006. The model organism as a system integrating "omics" data sets. *Nat. Rev. Mol. Cell Biol.* 7: 901–909.

Kim, M-S, et al. 2014. A draft map of the human proteome. *Nature* 509: 575–581.

Lander, E. S. 2011. Initial impact of sequencing the human genome. *Nature* 470: 187–197.

Lek, M., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536: 285–291, 2016.

Lupski, J. R. 2013. Genome mosaicism—One human, multiple genomes. *Science* 341: 358–359.

Mallick, S., et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538: 201–206.

Nezvizhskii, A. I., Vitek, O., and Aebersold, R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* 4: 787–797.

Noonan, J. P., et al. 2006. Sequencing and analysis of Neanderthal DNA. *Science* 314: 1113–1118.

Pennisi, E. 2006. The dawn of Stone Age genomics. *Science* 314: 1068–1071.

Qin, J., et al. 2010. A human gun microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–64.

Qin, N., et al. 2014. Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513: 59–64.

Reinert, K., Langmead, B., Weese, D., and Evers, D. J. 2015. Alignment of next-generation sequencing reads. *Annu. Rev. Genomics Hum. Genet.* 16: 133–151.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316: 222–234.

Schweitzer, M. H., et al. 2007. Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science* 316: 277–280.

Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Stronglyocentrotus purpuratus*. *Science* 314: 941–956.

Switnoski, M., Szczerbal, I., and Nowacka, J. 2004. The dog genome map and its use in mammalian comparative genomics. *J. Appl. Physiol.* 45: 195–214.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 492: 56–65.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526: 68–74.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 57–74. Part of ENCODE: Guidebook to the Human Genome. [A special edition of *Nature*]. *Nature* 489: 45–113.

The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–214.

____. 2012. A framework for human microbiome research. *Nature* 486: 215–221.

Thieman, W. J., and Palladino, M. A. 2013. *Introduction to biotechnology,* 3rd ed. San Francisco, CA: Pearson Education, Inc.

Venter, J. C., et al. 2001. The sequence of the human genome. *Science* 291: 1304–1351.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., and Watson, M. 2015. Exome sequencing: Current and future perspectives. *Genes Genomes Genetics* 5: 1543–1550.

Weinstock, G. M. 2012. Genomic approaches to studying the human microbiota. *Nature* 489: 250–256.

Wilhelm, M., et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509: 582–587.

CHAPTER 22 Applications of Genetic Engineering and Biotechnology

Altshuler, D., Daly, M. J., and Lander, E. S. 2008. Genetic mapping in human disease. *Science* 322: 881–888.

Anower, J., Wagner, S., McCracken, D., Wells, N., and Laible, G. 2012. Targeted microRNA expression in dairy cattle directs production of beta-lactoglobulin-free, high-casein milk. *Proc. Natl. Acad. Sci.* (*USA*) [it] 109: 16811–16816.

Denoeud, F., et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345: 1181–1184.

Dey, S. S., et al. 2015. Integrated genome and transcriptome sequencing of the same cell. *Nature* Biotech. 33: 285–289.

Dykxhoorn, D. M., and Liberman, J. 2006. Knocking down disease with siRNAs. *Cell* 126: 231–235.

Engler, O. B., et al. 2001. Peptide vaccines against hepatitis B virus: From animal model to human studies. *Mol. Immunol.* 38: 457–465.

Fan, H. C., et al. 2012. Non-invasive prenatal measurement of the fetal genome. *Nature* 487: 320–324.

Friend, S. H., and Stoughton, R. B. 2002. The magic of microarrays. *Sci. Am.* (Feb.) 286: 44–50.

Gibson, D. G., et al. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329: 52–56.

Green, E. D., and Guyer, M. S. 2011. Charting a course for genomic medicine from base pairs to bedside. *Nature* 470: 204–213.

Grubaugh, N.D., et al. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546: 401–405.

Hutchinson, C.A., III, et al. 2016. Design and synthesis of a minimal bacterial genome. *Science* 351: 1414.

Jiang, Y. H., et al. 2013. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* 93: 249–263.

Knoppers, B. M., Bordet, S., and Isasi, R. M. 2006. Preimplantation genetic diagnosis: An overview of socio-ethical and legal considerations. *Annu. Rev. Genom. Hum. Genet.* 7: 201–221.

Lajoie, M. J., Kosuri, S., Mosberg, J. A., Gregg, C. J., Zhang, D., and Church, G. M. 2013. Probing the limits of genetic recoding in essential genes. *Science* 342: 361–363.

Lajoie, M. J., et al. 2013. Genomically recoding organisms expand biological functions. *Science* 342: 357–360.

Lipsitch, M., and Inglesby, T. V. 2014. Moratorium on research intended to create novel potential pandemic pathogens. *mBio* 5: 1–6.

Lu, S., et al. 2012. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338: 1627–1630.

Maher, B. 2011. Human genetics: Genomes on prescription. *Nature* 478: 22–24.

Manolio, T. A., et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461: 747–753.

Manolio, T. A. 2017. A decade of shared genomic associations. *Nature* 546: 360–361.

Ostrov, M. et al. 2016. Design, synthesis, and testing toward a 57-codon genome. *Science* 353: 819–822.

Schizophrenia Working Group of the Psychiatric Genomes Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511: 421–427.

Shastry, B. S. 2006. Pharmacogenetics and the concept of individualized medicine. *Pharmacogenetics J.* 6: 16–21.

Siuti, P., Yazbek, J., and Lu, T. K. 2013. Synthetic circuits integrating logic and memory in living cells. *Nature Biotech.* 31: 448–452.

Stubbington, M.J.T. et al. 2017. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358: 58–64.

Wang, D. G., et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280: 1077–1082.

Wang, J., et al. 2012. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150: 402–412.

Wang, T., et al. 2016. Identification and characterization of essential genes in the human genome. *Science* 350: 1096–1101.

Whitelaw, C. B. A. 2004. Transgenic livestock made easy. *Trends Biotechnol.* 22(4): 257–259.

Wright, A.V., et al. 2016. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* 164: 29–44.

CHAPTER 23 Developmental Genetics

Anderson, E. R., Sandberg, R., and Lendahl, U. 2011. Notch signaling: Simplicity in design, versatility in function. *Development* 138: 3593–3612.

Cantone, I., and Fisher, A. G. 2013. Epigenetic programming and reprogramming during development. *Nat. Struct. Mol. Biol.* 20: 282–289.

Davidson, E. H., and Levine, M. S. 2008. Properties of developmental gene regulatory networks. *Proc. Nat. Acad. Sci.* 115: 20063–20066.

De Leon, S. B-T., and Davidson, E. H. 2007. Gene regulation: Gene control networks in development. *Annu. Rev. Biophys. Biomol. Struct.* 36: 191–212.

Eklund, E. 2011. The role of Hox proteins in leukemogenesis: Insights into key regulatory events in hematopoiesis. *Crit. Rev. Oncol.* 16: 65–76.

Feng, S., Jacobsen, S. E., and Reik, W. 2010. Epigenetic reprogramming in plant and animal development. *Science* 330: 622–627.

Gehring, W. J. 2012. The animal body plan, the prototypic body segment, and eye evolution. *Evol. Dev*. 14: 34–46.

Goodman, F. 2002. Limb malformations and the human *HOX* genes. *Am. J. Med. Genet*. 112: 256–265.

Gramzow, L., and Theissen, G. 2010. A hitchhiker's guide to the MADS world of plants. *Genome Biol*. 11: 214.

Gridley, T. 2003. Notch signaling and inherited human diseases. *Hum. Mol. Genet*. 12: R9–R13.

Inoue, T., et al. 2005. Gene regulatory special feature: Transcriptional network underlying *Caenorhabditis elegans* vulval development. *Proc. Nat. Acad. Sci*. 102: 4972–4977.

Kestler, H. A., Wawra, C., Kracher, B., and Kuhl, M. 2008. Network modeling of signal transduction: Establishing the global view. *Bioessays* 30: 1110–1125.

Krizek, B. A., and Fletcher, J. C. 2006. Molecular mechanisms of flower development: An armchair guide. *Nat. Rev. Genet*. 6: 688–698.

Lynch, J. A., and Roth, S. 2011. The evolution of dorsal-ventral patterning mechanism in insects. *Genes Dev*. 25: 107–118.

Maeda, R. K., and Karch, F. 2006. The ABC of the BX-C: The bithorax complex explained. *Development* 133: 1413–1422.

Nüsslein-Volhard, C., and Weischaus, E. 1980. Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287: 795–801.

Quinonez, S. C., and Innis, J. W. 2014. Human *HOX* disorders. *Mol. Genet. Metabol*. 111: 4–15.

Rebeiz, M., Patel, N. H., and Hinman, V. F. 2015. Unraveling the tangled skein: The evolution of transcriptional networks in development. *Annu. Rev. Genomics Hum. Genet*. 16: 103–131.

Reyes, J. C. 2006. Chromatin modifiers that control plant development. *Curr. Opin. Plant Biol*. 9: 21–27.

Schroeder, M. D., et al. 2004. Transcriptional control in the segmentation gene network of *Drosophila*. *PLOS Biol*. 2: 1396–1410.

Schvartsman, S. Y., Coppey, M., and Berezhkovskii, A. 2008. Dynamics of maternal morphogen gradients in *Drosophila*. *Curr. Opin. Genet. Dev*. 18: 342–347.

Verakasa, A., Del Campo, M., and McGinnis, W. 2000. Developmental patterning genes and their conserved functions: From model organisms to humans. *Mol. Genet. Metabol*. 69: 85–100.

Walser, C. B., and Lipshitz, H. D. 2011. Transcript clearance during the maternal-to-zygote transition. *Curr. Opin. Genet. Dev*. 21: 431–443.

Wang, M., and Sternberg, P. W. 2001. Pattern formation during *C. elegans* vulval induction. *Curr. Top. Dev. Biol*. 51: 189–220.

Yuan, J., and Kroemer, G. 2010. Alternative cell death mechanisms in development and beyond. *Genes Dev*. 24: 2592–2602.

CHAPTER 24 Cancer Genetics

Alison, M. R., et al. 2010. Stem cells in cancer: Instigators and propagators? *J. Cell Sci.* 123: 2357–2368.

Ashworth, A., and Hudson, T. J. 2013. Comparisons across cancers. *Nature* 502: 306–307.

Bernards, R., and Weinberg, R. A. 2002. A progression puzzle. *Nature* 418: 823.

Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Cancer* 8: 286–297.

Gray, J. 2010. Genomics of metastasis. *Nature* 464: 989–990.

Hartwell, L. H., and Kastan, M. B. 1994. Cell cycle control and cancer. *Science* 266: 1821–1827.

Kandoth, C., et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* 502: 333–340.

Marte, B. 2013. Tumour heterogeneity. *Nature* 501: 327.

Martincorena, I., and Campbell, P. J. 2016. Somatic mutation in cancer and normal cells. *Science* 349: 1483–1489.

Nurse, P. 1997. Checkpoint pathways come of age. *Cell* 91: 865–867.

Varmus, H. 2006. The new era in cancer research. *Science* 312: 1162–1164.

CHAPTER 25 Quantitative Genetics and Multifactorial Traits

Browman, K. W. 2001. Review of statistical methods of QTL mapping in experimental crosses. *Lab Animal* 30: 44–52.

Cong, B., Barrero, L. S., and Tanksley, S. D. 2008. Regulatory changes in a YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* 40: 800–804.

Cong, B., Liu, J., and Tanksley, S. 2002. Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proc. Nat. Acad. Sci.* 99: 13606–13611.

Crow, J. F. 1993. Francis Galton: Count and measure, measure and count. *Genetics* 135: 1.

Druka, A., et al. 2010. Expression of quantitative trait loci analysis in plants. *Plant Biotech.* 8: 10–27.

Falconer, D. S., and Mackay, F. C. 1996. *Introduction to quantitative genetics,* 4th ed. Essex, UK: Longman.

Farber, S. 1980. *Identical twins reared apart.* New York: Basic Books.

Feldman, M. W., and Lewontin, R. C. 1975. The heritability hangup. *Science* 190: 1163–1166.

Frary, A., et al. 2000. fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289: 85–88.

Gupta, V., et al. 2009. Genome analysis and genetic enhancement of tomato. *Crit. Rev. Biotech*. 29: 152–181.

Haley, C. 1991. Use of DNA fingerprints for the detection of major genes for quantitative traits in domestic species. *Anim. Genet*. 22: 259–277.

____. 1996. Livestock QTLs: Bringing home the bacon. *Trends Genet*. 11: 488–490.

Johnson, W. E., et al. 2010. Genetic restoration of the Florida panther. *Science* 329: 1641–1644.

Kaminsky, A. A., et al. 2009. DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* 41: 240–245.

Lander, E., and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185–199.

Lander, E., and Schork, N. 1994. Genetic dissection of complex traits. *Science* 265: 2037–2048.

Lynch, M., and Walsh, B. 1998. *Genetics and analysis of quantitative traits.* Sunderland, MA: Sinauer Associates.

Macy, T. F. C. 2001. Quantitative trait loci in *Drosophila*. *Nat. Rev. Genetics* 2: 11–19.

Newman, H. H., Freeman, F. N., and Holzinger, K. T. 1937. *Twins: A study of heredity and environment.* Chicago: University of Chicago Press.

Paterson, A., Deverna, J., Lanini, B., and Tanksley, S. 1990. Fine mapping of quantitative traits loci using selected overlapping recombinant chromosomes in an interspecific cross of tomato. *Genetics* 124: 735–742.

Tanksley, S. D. 2004. The genetic, developmental and molecular bases of fruit size and shape variation in tomato. *Plant Cell* 16: S181–S189.

Veenma, D., et al. 2012. Copy number detection in discordant monozygotic twins of congenital diaphragmatic hernia (CDH) and esophageal atresia (EA). *Eur. J. Hum. Genet.* 20: 298–306.

Zar, J. H. 1999. *Biostatistical analysis,* 4th ed. Upper Saddle River, NJ: Prentice Hall.

CHAPTER 26 Population and Evolutionary Genetics

Ansari-Lari, M. A., et al. 1997. The extent of genetic variation in the *CCR5* gene. *Nature Genetics* 16: 221–222.

Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M., and Meyer, A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* 439: 719–723.

Green, R. E., et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722.

Hedges, S. B., and Kumar, S. 2003. Genomic clocks and evolutionary time scales. *Trends Genet.* 19: 200–206.

Karn, M. N., and Penrose, L. S. 1951. Birth weight and gestation time in relation to maternal age, parity and infant survival. *Ann. Eugen.* 16: 147–164.

Kerr, W. E., and Wright, S. 1954. Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila melanogaster*. *Evolution* 8: 172–177.

Knowlton, N., et al. 1993. Divergence in proteins, mitochondrial DNA, and reproductive compatibility across the Isthmus of Panama. *Science* 260: 1629–1632.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature* 304: 412–417.

Lamb, R. S., and Irish, V. F. 2003. Functional divergence within the *APETALA3/PISTILLATA* floral homeotic gene lineages. *Proc. Natl. Acad. Sci.* 100: 6558–6563.

Leibert, F., et al. 1998. The *DCCR5* mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in northeastern Europe. *Hum. Mol. Genet.* 7: 399–406.

Meyer, M., Kircher, J., Gansauge, M. T., et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338: 222–226.

Mowat, A. 2017. Why does cystic fibrosis display the prevalence and distribution observed in human populations? *Curr. Pediatr. Res.* 21: 164–171.

Nei, M., Suzuki, Y., and Nozwa, M. 2010. The neutral theory of molecular evolution in the era of genomics. *Annu. Rev. Genomics Hum. Genet.* 11: 265–289.

Nielsen, R., et al. 2017. Tracing the peopling of the world through genomics. *Nature* 541: 302–310.

Noonan, J. P., et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314: 1113–1118.

Pagni, L., Lawson, D. J., Jagoda, E., et al. 2016. Genomic analyses inform on migration events in the peopling of Eurasia. *Nature* 538: 238–242.

Presgraves, D. C. 2010. The molecular evolutionary basis for species formation. *Nat. Rev. Genet.* 11: 175–180.

Reznick, D. N., and Ricklefs, R. E. 2009. Darwin's bridge between microevolution and macroevolution. *Nature* 457: 837–842.

Salzburger, W. 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Mol. Evol*. 18: 169–185.

Sankararaman, S., et al. 2016. The combined landscape of Denisovan and Neaderthal ancestry in present-day humans. *Curr. Biol*. 26: 1–7.

Schulter, D. 2009. Evidence for ecological speciation and its alternative. *Science* 323: 737–741.

Stiassny, M. L. J., and Meyer, A. 1999. Cichlids of the Rift Lakes. *Sci. Am.* (Feb.) 280: 64–69.

Vitti, J. J., Cho, M. K., Tishkoff, S. A., and Sabeti, P. C. 2012. Human evolutionary genetics: Ethical and interpretative issues. *Trends Genet*. 28: 137–145.

Yi, Z., et al. 2003. A 122.5 kilobase deletion of the *P* gene underlies the high prevalence of oculocutaneous albinism type 2 in the Navajo population. *Am. J. Hum. Genet*. 72: 62–72.

This page intentionally left blank

Appendix B

Answers to Selected Problems

CHAPTER 1

2. Your essay should include a description of the impact of recombinant DNA technology on the following: plant and animal husbandry and production, drug development, medical advances, forensics, and understanding gene function.

4. The genotype of an organism is defined as the specific allelic or genetic constitution of an organism, or, often, the allelic composition of one or a limited number of genes under investigation. The observable feature of those genes is called the phenotype. A gene variant is called an allele. There can be many such variants in a population, but for a diploid organism, only two such alleles can exist in any given individual.

6. *Genes* are the functional units of heredity. They consist of linear sequences of nucleotides and usually exert their influence by producing proteins through the processes of transcription and translation. *Chromosomes* are long strands of nucleotides that contain linear assemblies of genes. In many organisms, they associate with specific proteins. Chromosomes (and by extension, genes) are duplicated by a variety of enzymes so that daughter cells inherit copies of the parental hereditary information.

8. The central dogma of molecular genetics refers to the relationships among DNA, RNA, and proteins. The processes of transcription and translation are integral to understanding these relationships. Because DNA and RNA are discrete chemical entities, they can be isolated, studied, and manipulated in a variety of experiments that define modern genetics.

10. Restriction enzymes (endonucleases) cut double-stranded DNA at particular base sequences. When a vector is cleaved with the same enzyme, complementary ends are created such that ends, regardless of their origin, can be combined and ligated to form intact double-stranded structures. Such recombinant forms are often useful for industrial, research, and/or pharmaceutical efforts.

12. Unique transgenic plants and animals can be patented, as ruled by the U.S. Supreme Court in 1980. Supporters of organismic patenting argue that it is needed to encourage innovation and allow the costs of discovery to be recovered. Capital investors assume that there is a likely chance that their investments will yield positive returns. Others argue that natural substances should not be privately owned and that once they are owned by a small number of companies, free enterprise will be stifled.

14. Model organisms are not only useful, but also necessary for understanding genes that influence human diseases. Given that many genetic/molecular systems are highly conserved across broad phylogenetic lines, what is learned in one organism is usually applied to all organisms. In addition, most model organisms have peculiarities, such as ease of growth, genetic understanding, or abundant offspring, that make their use straightforward and especially informative in genetic studies.

16. For approximately 60 years discoveries in genetics have guided our understanding of living systems, aided rational drug design, and dominated many social discussions. Genetics provides the framework for universal biological processes and helps explain species stability and diversity. Given the central focus of genetics in so many of life's processes, it is understandable why so many genetic scientists have been awarded the Nobel Prize.

CHAPTER 2

Answers to Now Solve This

2.1 (a) 32, (b) 16

2.2 (a) 8, (b), 8 (c) 8

2.3 Not necessarily; if crossing over occurs in meiosis I, then the chromatids in the secondary oocyte are not identical. Once they separate during meiosis II, unlike chromatids, they reside in the ootid and the second polar body.

Solutions to Problems and Discussion Questions

2. Mitosis maintains chromosomal constancy, so there is no change in chromosome number or kind in the two daughter cells. By contrast, meiosis provides for a reduction in chromosome number and an opportunity for exchange of genetic material between homologous chromosomes. This leads to the production of numerous potentially different haploid (n) cells. During oogenesis, only one of the four meiotic products is functional; however, four of the four meiotic products of spermatogenesis are potentially functional. Errors during either mitosis or meiosis (such as nondisjunction events) can lead to cells with too many or too few chromosomes.

4. Chromosomes that are homologous share many properties including *overall length, position of the centromere (metacentric, submetacentric, acrocentric, telocentric), banding patterns, type and location of genes, and autoradiographic pattern. Diploidy* is a term often used in conjunction with the symbol $2n$. It means that both members of a homologous pair of chromosomes are present. *Haploidy* refers to the presence of a single copy of each homologous chromosome (n).

8. The mechanism of cytokinesis differs between these two cell types. In plants, a cell plate that was laid down during telophase becomes the middle lamella where primary and secondary layers of the cell wall are deposited. In animals, constriction of a cell membrane produces a cell furrow of daughter cells.

10. (a) *Synapsis* is the point-by-point pairing of homologous chromosomes during prophase of meiosis I.

(b) *Bivalents* are those structures formed by the synapsis of homologous chromosomes. In other words, there are two chromosomes (and four chromatids) that make up a bivalent.

(c) *Chiasmata* is the plural form of chiasma and refers to the structure, when viewed microscopically, of crossed chromatids.

(d) *Crossing over* is the exchange of genetic material between chromatids. It is a method of providing genetic variation through the breaking and rejoining of chromatids.

(e) *Chromomeres* are bands of chromatin that look different from neighboring patches along the length of a chromosome.

(f) *Sister chromatids* are "post-S phase" structures of replicated chromosomes. Sister chromatids are genetically identical (except where mutations have occurred) and are originally attached to the same centromere.

(g) *Tetrads* are synapsed homologous chromosomes thereby composed of four chromatids.

(h) *Dyads* are composed of two chromatids joined by a centromere.

(i) At anaphase II of meiosis, the centromeres divide and sister chromatids (*monads*) go to opposite poles.

12. During meiosis I, chromosome number is reduced to haploid complements. This is achieved by synapsis of homologous chromosomes and their subsequent separation. It would seem to be more mechanically difficult for genetically identical daughters to form from mitosis if homologous chromosomes paired. By having chromosomes unpaired at metaphase of mitosis, only centromere division is required for daughter cells to eventually receive identical chromosomal complements.

14. First, through independent assortment of chromosomes at anaphase I of meiosis, daughter cells (secondary spermatocytes and secondary oocytes) may contain different sets of maternally and paternally derived chromosomes. Second, crossing over, which happens at a much higher frequency in meiotic cells as compared with mitotic cells, allows maternally and paternally derived chromosomes to exchange segments, thereby increasing the likelihood that daughter cells (that is, secondary spermatocytes and secondary oocytes) are genetically unique. By contrast, daughter cells resulting from mitosis are usually genetically identical.

16. There would be 16 combinations with the addition of another chromosome pair.

18. One-half of each tetrad will have a maternal homolog: $(1/2)^{10}$.

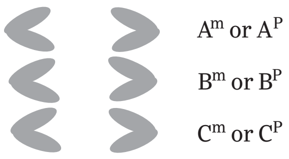
20. In angiosperms, meiosis results in the formation of microspores (male) and megaspores (female), which give rise to the haploid male and female gametophyte stage. Micro- and megagametophytes produce the pollen and the ovules, respectively. Following fertilization, the sporophyte is formed.

22. The folded-fiber model is based on each chromatid consisting of a single fiber wound like a skein of yarn. Each fiber consists of DNA and protein. A coiling process occurs during the transition of interphase chromatin into more condensed chromosomes during prophase of mitosis or meiosis. Such condensation leads to a 5000-fold contraction in the length of the DNA within each chromatid.

24. 50, 50, 50, 100, 200

26. At the end of prophase I, maternal and paternal copies of each homologous chromosome (A^m and A^p, B^m and B^p, C^m and C^p) will be synapsed. At the completion of anaphase I, eight possible combinations of products (A^m or A^p, B^m or B^p, C^m or C^p) will occur.

28. Eight (2 × 2 × 2) combinations are possible.



30. Fertilization of the gametes described in Problem 29 will give the following zygotes:

Zygote 1:	two copies of chromosome A
	two copies of chromosome B
	three copies of chromosome C
Zygote 2:	two copies of chromosome A
	two copies of chromosome B
	one copy of chromosome C

None of the zygotes will be diploid.

32. (Assume a normal chromosomal composition of the mother.) The secondary oocyte would completely lack chromosome 21. The resulting zygote would have one copy of chromosome 21 (from the father) and two copies of all the other chromosomes.

34. The secondary oocyte would have both a dyad and a monad from chromosome 21. Depending on how the monad partitioned at meiosis II, you would have either a normal chromosome 21 complement (the zygote that did not receive the monad) or a chromosome 21 trisomy in which the zygote received two number 21 chromosomes from the mother and one from the father.

CHAPTER 3

Answers to Now Solve This

3.1 P = checkered; p = plain

Cross: $PP \times PP$ or $PP \times Pp$

Notice in cross (d) that the checkered offspring, when crossed to plain, produce only checkered F_2 progeny and in cross (g) when crossed to checkered still produce only checkered progeny. From this additional information, one can conclude that in the progeny of cross (a), there are no heterozygotes and the original cross must have been $PP \times PP$.

Cross (b): $PP \times pp$

Cross (c): Because all the offspring from this cross are plain, there is no doubt that the genotype of both parents is *pp*.

Genotypes of all individuals:

	Progeny	
P_1 *Cross*	*Checkered*	*Plain*
(a) *PP* × *PP*	*PP*	
(b) *PP* × *pp*	*Pp*	
(c) *pp* × *pp*		pp
(d) *PP* × *pp*	Pp	
(e) *Pp* × *pp*	Pp	pp
(f) *Pp* × *Pp*	PP, Pp	pp
(g) *PP* × *Pp*	PP, Pp	

3.2 Suggested symbolism:

w = wrinkled seeds g = green cotyledons
W = round seeds G = yellow cotyledons

(a) Notice a 3 : 1 ratio for seed shape, therefore *Ww* × *Ww*; and no green cotyledons, therefore *GG* × *GG* or *GG* × *Gg*. Putting the two characteristics together gives

WwGG × *WwGG* or *WwGG* × *WwGg*

(b) *wwGg* × *WwGg*
(c) *WwGg* × *WwGg*
(d) *WwGg* × *wwgg*

3.3 (a)

Offspring:

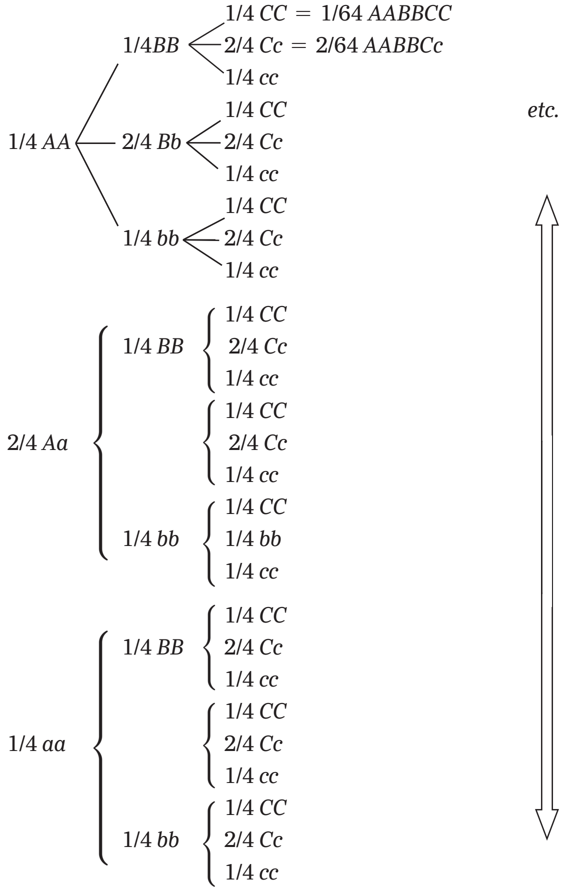
Genotypes	*Ratio*	*Phenotypes*
AABBCC	(1/16)	*A_B_C_* = 12/16
AABBCc	(1/16)	
AABbCC	(1/16)	
AABbCc	(1/16)	
AaBBCC	(2/16)	
AaBBCc	(2/16)	
AaBbCC	(2/16)	
AaBbCc	(2/16)	
aaBBCC	(1/16)	*aaB_C_* = 4/16
aaBBCc	(1/16)	
aaBbCC	(1/16)	
aaBbCc	(1/16)	

(b)

Offspring:

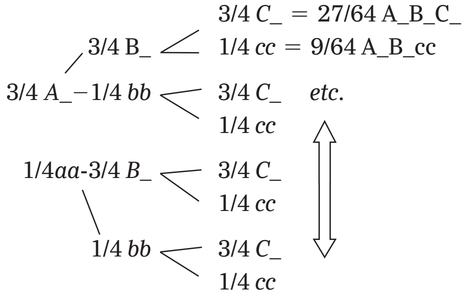
Genotypes	*Ratio*	*Phenotypes*
AaBBCC	1/8	*A_BBC_* = 3/8
AaBBCc	2/8	
AaBBcc	1/8	*A_BBcc* = 1/8
aaBBCC	1/8	*aaBBC_* = 3/8
aaBBCc	2/8	
aaBBcc	1/8	*aaBBcc* = 1/8

(c) There will be eight (2^n) different kinds of gametes from each of the parents and therefore a 64-box Punnett square. Doing this problem by the forked-line method helps considerably.



Simply multiply through each component to arrive at the final genotypic frequencies.

For the phenotypic frequencies, set up the problem in the following manner:



3.4 One must think of this problem as a dihybrid F_2 situation with the following expectations:

Expected ratio	*Observed (o)*	*Expected (e)*
9/16	315	312.75
3/16	108	104.25
3/16	101	104.25
1/16	32	34.75

$\chi^2 = 0.47$

Looking at the table in the text, one can see that this value is associated with a probability greater than 0.90 for 3 degrees of freedom (because there are now four classes in the test). The observed and expected values do not deviate significantly.

To deal with parts (b) and (c) it is easier to see the observed values for the monohybrid ratios if the phenotypes are listed:

smooth, yellow	315
smooth, green	108
wrinkled, yellow	101
wrinkled, green	32

For the smooth: wrinkled *monohybrid component*, the smooth types total 423 (315 + 108), while the wrinkled types total 133 (101 + 32).

Expected ratio	*Observed (o)*	*Expected (e)*
3/4	423	417
1/4	133	139

The value is 0.35, and in examining the text for 1 degree of freedom, the *p* value is greater than 0.50 and less than 0.90. We fail to reject the null hypothesis and are confident that the observed values do not differ significantly from the expected values.

(c) For the yellow: green portion of the problem, see that there are 416 yellow plants (315 + 101) and 140 (108 + 32) green plants.

Expected ratio	*Observed (o)*	*Expected (e)*
3/4	416	417
1/4	140	139

The value is 0.01, and in examining the text for 1 degree of freedom, the *p* value is greater than 0.90. We fail to reject the null hypothesis and are confident that the observed values do not differ significantly from the expected values.

3.5 The gene is inherited as an autosomal recessive. Notice that two normal individuals II-3 and II-4 have produced a daughter (III-2) with myopia.
I-1 (*aa*), I-2 (*Aa* or *AA*), I-3 (*Aa*), I-4 (*Aa*)
II-1 (*Aa*), II-2 (*Aa*), II-3 (*Aa*), II-4 (*Aa*), II-5 (*aa*), II-6 (*AA* or *Aa*), II-7 (*AA* or *Aa*)
III-1 (*AA* or *Aa*), III-2 (*aa*), III-3 (*AA* or *Aa*)

Solutions to Problems and Discussion Questions

2. Your essay should include the following points: 1. Factors occur in pairs. 2. Some genes have dominant and recessive alleles. 3. Alleles segregate from each other during gamete formation. When homologous chromosomes separate from each other at anaphase I, alleles will go to opposite poles of the meiotic apparatus. 4. One gene pair separates independently from other gene pairs. Different gene pairs on the same homologous pair of chromosomes (if far apart) or on nonhomologous chromosomes will separate independently from each other during meiosis.

4. Unit factors in pairs, dominance and recessiveness, segregation

6. *Pisum sativum* is easy to cultivate. It is naturally self-fertilizing, but it can be crossbred. It has several visible features (e.g., tall or short, red flowers or white flowers) that are consistent under a variety of environmental conditions, yet contrast due to genetic circumstances. Seeds could be obtained from local merchants.

8. *WWgg* = 1/16

10. Several points surface in the first sentence of this question. First, two alternatives (black and white) of one characteristic (coat color) are being described; therefore, a monohybrid condition exists. Second, which is dominant, *black* or *white*? Note that all the offspring are black; therefore, black can be considered dominant. The second sentence of the problem verifies that a monohybrid cross is involved because of the 3/4 black and 1/4 white distribution in the offspring. Referring to the appropriate figures and knowing that genes occur in pairs in diploid organisms, one can write the genotypes and the phenotypes requested in part (a) as follows:

(a)

P_1:			
Phenotypes:	black	×	white
Genotypes:	*WW*		*ww*
Gametes:	Ⓦ		ⓦ
F_1:	*Ww* (black)		

F_1 × F_1:				
Phenotypes:	black	×	black	
Genotypes:	*Ww*		*Ww*	
Gametes:	Ⓦ	ⓦ	Ⓦ	ⓦ
(combine as in the text)				
F_2:				
Phenotypes:	black	black	black	white
Genotypes:	*WW*	*Ww*	*Ww*	*ww*

(b) Since *white* is a recessive gene (to *black*), each white guinea pig must be homozygous and a cross between two white guinea pigs must produce all white offspring.

white	×	white
ww		*ww*

(c) Recall the various possibilities of the genotypes capable of producing the black phenotype in the F_2 generation in part (a) above: *WW* and *Ww*. In Cross 1 in the problem, all black offspring are observed and the most likely parental genotypes would be as follows:

WW	×	*WW*
	or	
WW	×	*Ww*

There is the possibility that black guinea pigs of the *Ww* genotype could produce all black offspring if the sample size was such that *ww* offspring were not produced. In Cross 2, a typical 3:1 Mendelian ratio is observed, which indicates that two heterozygotes were crossed:

$$\boxed{Ww \quad \times \quad Ww}$$

12. There are two characteristics presented here: body color and wing length. First, assign meaningful gene symbols.

Body color	*Wing length*
E = gray body color	*V* = long wings
e = ebony body color	*v* = vestigial wings

(a) P_1:

EEVV × *eevv*

F_1: *EeVv* (gray, long)

F_2: This will be the result of a Punnett square with 16 boxes.

Phenotype	*Ratio*	*Genotype*	*Ratio*
gray, long	9/16	*EEVV*	1/16
		EEVv	2/16
		EeVV	2/16
		EeVv	4/16
gray, vestigial	3/16	*EEvv*	1/16
		Eevv	2/16
ebony, long	3/16	*eeVV*	1/16
		eeVv	2/16
ebony, vestigial	1/16	*eevv*	1/16

(b) P_1: *EEvv* × *eeVV*

F_1: It is important to see that the results from this cross will be exactly the same as those in part (a) above. The only difference is that the recessive genes are coming from both parents, rather than from one parent only as in part (a). The F_2 ratio will be the same as part (a) also. When you have genes on the autosomes (not X-linked), independent assortment, complete dominance, and no gene interaction (see later) in a cross involving double heterozygotes, the offspring ratio will be 9:3:3:1.

(c) P_1: *EEVV* × *EEvv*

F_1: *EEVv* (gray, long)

F_2: Notice that all the offspring will have gray bodies and you will get a 3:1 ratio of long to vestigial wings. You should see this before you even begin working through the problem. Even though this cross involves two gene pairs, it will give a "monohybrid" type of ratio because one of the gene pairs is homozygous (body color) and **one** gene pair is heterozygous (wing length).

Phenotype	*Ratio*	*Genotype*	*Ratio*
gray, long	3/4	*EEVV*	1/4
		EEVv	2/4
gray, vestigial	1/4	*EEvv*	1/4

14.

Phenotype	*Genotype*
P_1: yellow × green	*GG* × *gg*
F_1: all yellow	*Gg*
F_2: 6022 yellow	1/4 *GG*; 2/4 *Gg*
2001 green	1/4 *gg*

GG × *GG* = all *GG*
Gg × *Gg* = 1/4 *GG*; 2/4 *Gg*; 1/4 *gg*

16.

Seed shape	*Seed color*
W = round	*G* = yellow
w = wrinkled	*g* = green

P_1: *WWgg* × *wwGG*

F_1: *WwGg* cross to *wwgg*
(which is a typical testcross)
The offspring will occur in a typical 1 : 1 : 1 : 1 as
1/4 *WwGg* (round, yellow)
1/4 *Wwgg* (round, green)
1/4 *wwGg* (wrinkled, yellow)
1/4 *wwgg* (wrinkled, green)

18. (a)

Expected ratio	*Observed (o)*	*Expected (e)*
3/4	882	885.75
1/4	299	295.25

$$\chi^2 = \Sigma(o\text{-}e)^2/e = 0.064$$

By looking at the χ^2 table with 1 degree of freedom (because there were two classes and therefore $n - 1$ or 1 degree of freedom), we find a probability (p) value between 0.9 and 0.5.

We would therefore say that there is a "good fit" between the observed and expected values.

(b)

Expected ratio	*Observed (o)*	*Expected (e)*
3/4	705	696.75
1/4	224	232.25

$$\chi^2 = 0.39$$

The p value in the table for 1 degree of freedom is still between 0.9 and 0.5; however, because the χ^2 value is larger in part (b), we should say that the deviations from expectation are greater. The deviation in each case can be attributed to chance.

20. Use of the $p = 0.10$ as the "critical" value for rejecting or failing to reject the null hypothesis instead of $p = 0.05$ would allow more null hypotheses to be rejected. As the critical p value is increased, it takes a smaller χ^2 value to cause rejection of the null hypothesis. It would take less difference between the expected and observed values to reject the null hypothesis; therefore, the stringency of failing to reject the null hypothesis is increased.

22. (a) There are two possibilities. Either the trait is dominant, in which case I-1 is heterozygous as are II-2 and II-3, or the trait is recessive, and I-1 is homozygous and

I-2 is heterozygous. Under the condition of recessiveness, both II-1 and II-4 would be heterozygous; II-2 and II-3 are homozygous.

(b) recessive: parents *Aa, Aa*

(c) recessive: parents *Aa, Aa*

(d) recessive or dominant, not sex-linked; if recessive, parents of the first cross are most likely *AA* and *aa* (the unaffected parent could also be *Aa*) and parents of the second cross are *Aa*, *aa*; if dominant, both sets of parents are *aa* and *Aa*.

24. Given the birth of the second affected child, it is highly likely that both parents were carriers for a recessive mutant gene causing Smith–Lemli–Opitz syndrome. Under that circumstance, there is a 25 percent chance that each of their children would be affected. The probability that two children of heterozygous parents would be affected would be $0.25 \times 0.25 = 0.0625$, or a little over 6 percent.

26. (a) Analyze each trait separately. In the first cross, notice that a 3 : 1 ratio exists for the spiny to smooth phenotypes, leading to the hypothesis that the *spiny* allele is dominant to *smooth*. Apply the same reasoning to the second cross, where there is a 3 : 1 ratio of *purple* to *white*. We would also predict that the *purple* allele is dominant to *white*.

(b) One could cross a homozygous purple, spiny plant to a homozygous white, smooth plant. The purple, spiny F_1 would support the hypothesis that *purple* is dominant to *white* and *spiny* is dominant to *smooth*. In the F_2, a 9 : 3 : 3 : 1 ratio would not only support the above hypothesis, but also indicate the independent inheritance and expression of the two traits.

Also, there are only three possibilities: both are heterozygous, neither is heterozygous, and at least one is heterozygous. You have already calculated the first two probabilities; the last is simply $1 - (1/12 + 6/12) = 5/12$.

28. (a) Consider that the data represent a 3:1 ratio based on the information given in the problem: *Ss* × *Ss*.

For set I the χ^2 value would be

$$(30 - 26.25)^2/26.25 + (5 - 8.75)^2/8.75 = 2.15$$

with *p* being between 0.2 and 0.05 so one would accept the null hypothesis of no significant difference between the expected and observed values.

For set II, the χ^2 value would be 21.43 and $p < 0.001$. One would reject the null hypothesis and assume a significant difference between the observed and expected values.

(b) Clearly, with an increase in sample size, a different conclusion is reached. In most cases, more confidence is gained as the sample size increases. In fact, most statisticians recommend that the expected values in each class should not be less than 10.

30. (a) For initially rare recessive genes to become established, rare crosses between heterozygotes would have to occur at a relatively high frequency.

(b) The likelihood of the establishment of albinism in whales with a rare dominant gene is probably higher than if the gene were recessive.

CHAPTER 4

Answers to Now Solve This

4.1 (a) Parents: sepia × cream

Cross: $c^kc^a \times c^dc^a \Downarrow$

2/4 sepia; 1/4 cream; 1/4 albino

(b) Parents: sepia × cream

($c^kc^d \times c^kc^d$ or $c^kc^d \times c^kc^a$); the cream parent could be c^dc^d or c^dc^a.

Crosses: $c^kc^a \times c^dc^d \Downarrow$

1/2 sepia; 1/2 cream (if parents are assumed to be homozygous)

or $c^kc^a \times c^dc^a \Downarrow$

1/2 sepia; 1/4 cream; 1/4 albino

(c) Parents: sepia × cream

Because the sepia guinea pig had two full color parents, which could be

$$Cc^k,\ Cc^d,\ \text{or}\ Cc^a$$

(not *CC* because sepia could not be produced), its genotype could be

$$c^kc^k,\ c^kc^d,\ \text{or}\ c^kc^a$$

Because the cream guinea pig had two sepia parents ($c^kc^d \times c^kc^d$ or $c^kc^d \times c^kc^a$), the cream parent could be c^dc^d or c^dc^a.

Crosses:

$c^kc^k \times c^dc^d \Longrightarrow$ all sepia
$c^kc^k \times c^dc^a \Longrightarrow$ all sepia
$c^kc^d \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^d \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/4 cream; 1/4 albino

(d) Parents: sepia × cream

Because the sepia parent had a full color parent and an albino parent ($Cc^k \times c^ac^a$), it must be c^kc^a. The cream parent had two full color parents that could be Cc^d or Cc^a; therefore, it could be c^dc^d or c^dc^a.

Crosses:

$c^kc^a \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/4 cream; 1/4 albino

4.2 A = pigment; a = pigmentless (colorless); B = purple; b = red

AaBb × *AaBb*

$$\Downarrow$$

*A_B_* = purple
A_bb = red
*aaB_* = colorless
aabb = colorless

4.3 For all three pedigrees, let *a* represent the mutant gene and *A* represent its normal allele.

(a) This pedigree is consistent with an X-linked recessive trait because the male would contribute an X chromosome carrying the *a* mutation to the *aa* daughter. The mother would have to be heterozygous *Aa*.

(b) This pedigree is consistent with an X-linked recessive trait because the mother could be *Aa* and transmit her *a* allele to her one son (*a*/Y) and her *A* allele to her other son.

(c) This pedigree is not consistent with an X-linked recessive mode of inheritance because the *aa* mother has an *A*/Y son.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of alleles that do not function independently of each other or that reduce the viability of a class of offspring. With multiple alleles, there are more than two alternatives of a given gene at a given locus.

4. $Pp \times Pp$ ↴

1/4 *PP* (**lethal**)
2/4 *Pp* (platinum)
1/4 *pp* (silver)

Therefore, the ratio of surviving foxes is 2/3 platinum, 1/3 silver. The *P* allele behaves as a recessive in terms of lethality (seen only in the homozygote) but as a dominant in terms of coat color (seen in the homozygote).

6.

Blood Group (Phenotype)	*Genotype(s)*
A	I^AI^A, I^Ai
B	I^BI^B, I^Bi
AB	I^AI^B
O	ii

I^A and I^B are codominant (notice the AB blood group), and each is dominant to *i*.

8. The only *blood type* that would exclude a male from being the father would be AB, because no *i* allele is present. Because many individuals in a population could have genotypes with the *i* allele, one could not prove that a particular male was the father by this method.

10. The simplest explanation is that the homozygous creeper genotype is lethal.

12. Three independently assorting characteristics are being dealt with: (1) flower color (incomplete dominance), (2) flower shape (dominant/recessive), and (3) plant height (dominant/recessive).

Establish appropriate gene symbols:

Flower color: RR = red; Rr = pink; rr = white

Flower shape: P = personate; p = peloric

Plant height: D = tall; d = dwarf

$RRPPDD \times rrppdd$ ↴

RrPpDd (pink, personate, tall)

Use *components* of the forked-line method as follows:

2/4 pink × 3/4 personate × 3/4 tall = 18/64

14. (a) This is a case of incomplete dominance in which, as shown in the third cross, the heterozygote (palomino) produces a typical 1 : 2 : 1 ratio.

$C^{ch}C^{ch}$ = chestnut
C^cC^c = cremello
$C^{ch}C^c$ = palomino

(b) The F_1 resulting from matings between cremello and chestnut horses would be expected to be all palomino. The F_2 would be expected to fall in a 1 : 2 : 1 ratio as in the third cross in part (a).

16. (a) In a cross of $AACC \times aacc$, the offspring are all *AaCc* (agouti) because the *C* allele allows pigment to be deposited in the hair, and when it is, it will be agouti. F_2 offspring would have the following "simplified" genotypes with the corresponding phenotypes:

*A_C_* = 9/16 (agouti)

A_cc = 3/16 (colorless because *cc* is epistatic to *A*)

*aaC_* = 3/16 (black)

aacc = 1/16 (colorless because *cc* is epistatic to *aa*)

The two colorless classes are phenotypically indistinguishable; therefore, the final ratio is 9 : 3 : 4.

(b) Results of crosses of female agouti (*A_C_*) × *aacc* (males) are given in three groups:
(1) To produce an even number of agouti and colorless offspring, the female parent must have been *AACc* so that half of the offspring are able to deposit pigment because of *C*; when they do, they are all agouti (having received only *A* from the female parent).
(2) To produce an even number of agouti and black offspring, the mother must have been *Aa*, and so that no colorless offspring were produced, the female must have been *CC*. Her genotype must have been *AaCC*.
(3) Notice that half of the offspring are colorless; therefore, the female must have been *Cc*. Half of the pigmented offspring are black and half are agouti; therefore, the female must have been *Aa*. Overall, the *AaCc* genotype seems appropriate.

18. (a) Since this is a 9 : 3 : 3 : 1 ratio with no albino phenotypes, the parents must each have been double heterozygotes and incapable of producing the *cc* genotype. Genotypes:

$AaBbCC \times AaBbCC$
or
$AaBbCC \times AaBbCc$

Phenotypes:

gray × gray

(b) Since there are no black offspring, there are no combinations in the parents that can produce *aa*. The 4/16 proportion indicates that the *C* locus is heterozygous in both parents.

If the parents are

$AABbCc \times AaBbCc$
or
$AABbCc \times AABbCc$

then the results would follow the pattern given.

Phenotypes: gray × gray

(c) Notice that 16/64 or 1/4 of the offspring are albino; therefore, the parents are both heterozygous at the *C* locus. Second, notice that without considering the *C* locus, there is a 27 : 9 : 9 : 3 ratio that reduces to a 9 : 3 : 3 : 1 ratio.

Given this information, the genotypes must be

$$AaBbCc \times AaBbCc$$

Phenotypes: gray × gray

(d) Genotypes: $aaBbCc \times aabbCc$
Phenotypes: black × cream

(e) Genotypes: $aaBbCc \times aaBbcc$
Phenotypes: black × albino

20. The initial cross must have been

$$AABB \times aabb$$

There are two gene pairs involved.

(a) $A\_B\_ = 9/16$ (tall)
$A\_bb = 3/16$ (dwarf)
$aaB\_ = 3/16$ (dwarf)
$aabb = 1/16$ (dwarf)

(b) To be true breeding, a plant needs to be homozygous at both loci. There are three different classes of dwarf plants. Within each of the 3/16 classes there are two types:

$A\_bb = 3/16$ (dwarf)
$= 1/3$ *AAbb* and 2/3 *Aabb*

and

$aaB\_ = 3/16$ (dwarf)
$= 1/3$ *aaBB* and 2/3 *aaBb*

Therefore, the true breeding dwarf plants would be the following:

AAbb, aaBB, and *aabb*

and they would constitute 3/7 of the dwarf group.
There is also one class of true breeding tall plant, *AABB*.
Taken together, these four classes constitute 1/4 of all plants.

22.

Cross 1 = (c)
Cross 2 = (d)
Cross 3 = (b)
Cross 4 = (e)
Cross 5 = (a)

24. The mating is $X^{RG}\ X^{rg}; I^{A}i^{*} \times X^{RG}\ Y; I^{A}i$.
The final product of the independent probabilities is

$$1/2 \times 1 \times 1/4 = 1/8$$

26. Assuming that the parents are homozygous, the crosses would be as follows. Notice that the *X* symbol may remain to remind us that the *sd* gene is on the X chromosome. It is extremely important that one account for both the mutant genes and each of their wild-type alleles.

P_1:$X^{sd}\ X^{sd}; e^{+}/e^{+} \times X^{+}/Y; e/e$

F_1:

1/2 $X^{+}X^{sd}; e^{+}/e$ (female, normal)
1/2 $X^{sd}/Y; e^{+}/e$ (male, scalloped)

F_2:

Phenotypes:
3/16 normal females
3/16 normal males
1/16 ebony females
1/16 ebony males
3/16 scalloped females
3/16 scalloped males
1/16 scalloped, ebony females
1/16 scalloped, ebony males

Forked-line method:

P_1: $X^{sd}\ X^{sd}; e^{+}/e^{+} \times X^{+}/Y; e/e$

$$\Downarrow$$

F_1: 1/2 $X^{+}X^{sd}; e^{+}/e$ (female, normal)
1/2 $X^{sd}/Y; e^{+}/e$ (male, scalloped)

F_2:

	Wings	Color	
1/4	females, normal	3/4 normal	— 3/16
		1/4 ebony	— 1/16
1/4	females, scalloped	3/4 normal	— 3/16
		1/4 ebony	— 1/16
1/4	males, normal	3/4 normal	— 3/16
		1/4 ebony	— 1/16
1/4	males, scalloped	3/4 normal	— 3/16
		1/4 ebony	— 1/16

28. (a) P_1:$X^{v}tX^{v}; +/+ \times X^{+}/Y; bw/bw$ ⬎

F_1:
1/2 $X^{+}\ X^{v}; +/bw$ (female, normal)
1/2 $X^{v}\ /Y; +/bw$ (male, vermilion)

F_2:
3/16 = females, normal
1/16 = females, brown eyes
3/16 = females, vermilion eyes
1/16 = females, white eyes
3/16 = males, normal
1/16 = males, brown eyes
3/16 = males, vermilion eyes
1/16 = males, white eyes

(b) P_1: $X^{+}\ X^{+}; bw/bw \times X^{v}/Y; +/+$ ⬎

F_1:
1/2 $X^{+}X^{v}; +/bw$ (female, normal)
1/2 $X^{+}/Y; +/bw$ (male, normal)

F_2:
6/16 = females, normal
2/16 = females, brown eyes
3/16 = males, normal
1/16 = males, brown eyes
3/16 = males, vermilion eyes
1/16 = males, white eyes

(c) P_1:$X^{v}\ X^{v}; bw/bw \times X^{+}/Y; +/+$ ⬎

F_1:1/2 $X^{+}\ X^{v}; +/bw$ (female, normal)
1/2 $X^{v}/Y; +/bw$ (male, vermilion)

F_2:

3/16 = females, normal
1/16 = females, brown eyes
3/16 = females, vermilion eyes
1/16 = females, white eyes
3/16 = males, normal
1/16 = males, brown eyes
3/16 = males, vermilion eyes
1/16 = males, white eyes

30. (a,b) The condition is *recessive*. In the second cross, note that the father is not shaded, yet the daughter (II-4) is. If the condition is recessive, then it must also be *autosomal*.

(c) II − 1 = *AA* or *Aa*
II − 6 = *AA* or *Aa*
II − 9 = *Aa*

32. P_1: female: *HH* × male: *hh*
F_1: all hen-feathering

F_2: *1/2 females*	*1/2 males*
1/4 *HH* hen-feathering	hen-feathering
2/4 *Hh* hen-feathering	hen-feathering
1/4 *hh* hen-feathering	cock-feathering

All the offspring would be hen-feathered except for 1/8 of the males, which are cock-feathered.

34. Phenotypic expression is dependent on the genome of the organism; the immediate molecular and cellular environment of the genome; and numerous interactions between a genome, the organism, and the environment.

36. (a) *AAB_* × *aaBB* (Other configurations are possible, but each must give all offspring with *A* and *B* dominant alleles.)
(b) *AaB_* × *aaBB* (Other configurations are possible, but no *bb* genotypes can be produced.)
(c) *AABb* × *aaBb*
(d) *AABB* × *aabb* (Other configurations are possible, but each must give all offspring with at least one *A* dominant allele.)
(e) *AaBb* × *Aabb*
(f) *AaBb* × *aabb*
(g) *aaBb* × *aaBb*
(h) *AaBb* × *AaBb*
Those genotypes that will breed true will be as follows:

black = *AABB*
golden = all genotypes that are *bb*
brown = *aaBB*

38. The homozygous dominant Dexter type is lethal. Polled is caused by an independently assorting dominant allele, while horned is caused by the recessive allele to polled.

40. Given the degree of outcrossing, that the gene is probably quite rare and therefore heterozygotes are uncommon, and that the frequency of transmission is high, it is likely that this form of male precocious puberty is caused by an autosomal dominant, sex-limited gene.

42. Given the following genotypes of the parents,

aabb = crimson
AABB = white

the F_1 consists of *AaBb* genotypes with a rose phenotype.

In the F_2 the following genotypes correspond to the given phenotypes:

AA__ = white 4/16 (any genotype at the *B* locus)
AaBB = magenta 2/16
AaBb = rose 4/16
Aabb = orange 2/16
aaBB = yellow 1/16
aaBb = pale yellow 2/16
aabb = crimson 1/16

Gene interaction is occurring along with the absence of complete dominance.

CHAPTER 5

Answers to Now Solve This

5.1 (a) 1/4 *AaBb*
1/4 *Aabb*
1/4 *aaBb*
1/4 *aabb*

(b) 1/4 *AaBb*
1/4 *Aabb*
1/4 *aaBb*
1/4 *aabb*

(c) If the arrangement is *AB*/*ab* × *ab*/*ab*
1/2 *AB*/*ab*
1/2 *ab*/*ab*
If the arrangement is *Ab*/*aB* × *ab*/*ab*
1/2 *Ab*/*ab*
1/2 *aB*/*ab*

5.2 Adding the crossover percentages together (6.9 + 7.1) gives 14 percent, which would be the map distance between the two genes. The dominant genes are on one homolog, while the recessive alleles are on the other homolog.

5.3 (a,b) + *b c*/*a* + +

$$a - b = \frac{32 + 38 + 0 + 0}{1000} \times 100 = 7 \text{ map units}$$

$$b - c = \frac{11 + 9 + 0 + 0}{1000} \times 100 = 2 \text{ map units}$$

(c) The progeny phenotypes that are missing are + + *c* and *a b* + , which, of 1000 offspring, 1.4 (0.07 × 0.02 × 1000) would be expected. Perhaps by chance or some other unknown selective factor, they were not observed.

Solutions to Problems and Discussion Questions

2. Your essay should include methods of detection through crosses with appropriate, distinguishable markers and that, in most cases, the frequency of crossing over is directly related to the distance between genes.

4. With some qualification, especially around the centromeres and telomeres, one can say that crossing over is somewhat randomly distributed over the length of the chromosome. Two loci that are far apart are more likely to have a crossover between them than two loci that are close together.

6. If the probability of one event is

$$1/X$$

then the probability of two events occurring at the same time will be

$$1/X^2$$

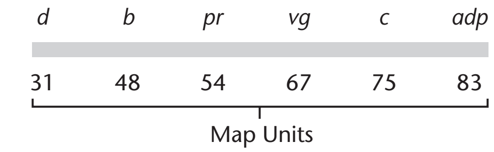
8. Each cross must be set up in such a way as to reveal crossovers because it is on the basis of crossover frequency that genetic maps are developed. It is necessary that genetic heterogeneity exists so that different arrangements of genes, generated by crossing over, can be distinguished. If crossing over in an organism is limited to one sex, the heterozygous individual must be of the sex in which this event occurs. In other words, it would be useless to map genes in *Drosophila* if the male parent is the heterozygote since crossing over is not typical in *Drosophila* males. Lastly, the cross must be set up so that the phenotypes of the offspring readily reveal their genotypes. The best arrangement is one where a fully heterozygous organism is crossed with an organism that is fully recessive for the genes being mapped.

10. Notice that the most frequent phenotypes in the offspring, the parentals, are colored, green (88) and colorless, yellow (92). This indicates that the heterozygous parent in the testcross is coupled

$$\mathrm{RY/ry} \times \mathrm{ry/ry}$$

with the two dominant genes on one chromosome and the two recessives on the homolog. Because there are 20 crossover progeny among the 200, or 20/200, the map distance would be 10 map units (20/200 × 100 to convert to percentages) between the *R* and *Y* loci.

12.



The expected map units between *d* and *c* would be 44, between *d* and *vg* would be 36, and between *d* and *adp* 52. However, because there is a theoretical maximum of 50 map units possible between two loci in any one cross, that distance would be below the 52 determined by simple subtraction.

14. (a) P$_1$:*sc sv*/*sc sv* × + + + /*Y*

F$_1$: + + + /*sc sv* × *sc sv*/*Y*

(b) Using method I or II for determining the sequence of genes, examine the parental classes and compare the arrangement with the double-crossover (least frequent) classes.

$$\frac{sc\ vs}{+\ +\ +}$$

$$\begin{aligned} sc - v &= \frac{150 + 156 + 10 + 14}{1000} \times 100 \\ &= 33, \text{(map units)} \\ v - s &= \frac{46 + 30 + 10 + 14}{1000} \times 100 \\ &= 10, \text{(map units)} \end{aligned}$$

$$\underbrace{sc\text{-------}v}_{33}\underbrace{\text{-------}s}_{10}$$

(c,d) The coefficient of coincidence = 0.727, which indicates that there were fewer double crossovers than expected; therefore, positive chromosomal interference is present.

16. (a) P$_1$:*D* + + / + + + × + *e p*/ + *e p*

F$_1$:*D* + + / + *e p* × + *e p*/ + *e p*

F$_2$:*D* + + / + *e p*	Dichaete
+ *e p*/ + *e p*	ebony, pink
D e + / + *e p*	Dichaete, ebony
+ + *p*/ + *e p*	pink
D + *p*/ + *e p*	Dichaete, pink
+ *e* + / + *e p*	ebony
D e p/ + *e p*	Dichaete, ebony, pink
+ + + / + *e p*	wild type

(b) Determine which gene is in the middle by comparing the parental classes with the double-crossover classes.

F$_1$: D + + / + p e × + p e/ + p e

$$\begin{aligned} D - p &= \frac{12 + 13 + 2 + 3}{1000} \times 100 \\ &= 3.0 \text{ map units} \\ p - e &= \frac{84 + 96 + 2 + 3}{1000} \times 100 \\ &= 18.5 \text{ map units} \end{aligned}$$

18. + *cu* / + *cu*

20. Because sister chromatids are genetically identical (with the exception of rare new mutations), crossing over between sisters provides no increase in genetic variability. Somatic crossing over would have no influence on the offspring produced.

22. (a) There would be $2^n = 8$ genotypic and phenotypic classes, and they would occur in a 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 ratio.

(b) There would be two classes, and they would occur in a 1 : 1 ratio.

(c) There are 20 map units between the *A* and *B* loci, and locus *C* assorts independently from both *A* and *B* loci.

24. In contrast to the other organisms mentioned, a single human mating pair produces relatively few offspring and the haploid number of chromosomes is relatively high (23), so there are rather small numbers of identifiable genes per chromosome. In addition, accurate medical records are often difficult to obtain, and the life cycle is relatively long.

26. Assign the following symbols:

R = red *r* = yellow
O = oval *o* = long

Progeny A: *R o*/*r O* × *o*/*r o* = 10 map units
Progeny B: *R O*/*r o* × *r o*/*r o* = 10 map units

28. Look for overlap between chromosome numbers in given clones and genes expressed. Note that *ENO1* is expressed in clones B, D, and E; chromosomes 1 and 5 are common to these clones. However, since *ENO1* is not expressed in clone C, which is missing chromosome 1 (and has chromosome 5), *ENO1* must be on chromosome 1.

MDH1: chromosome 2
PEPS: chromosome 4
PMG1: chromosome 1

30. There is no crossing over in males, so if two genes are on the same chromosome, there will be complete linkage of the genes in the male gametes. In females, crossing over will produce parental and crossover gametes. What you will have is the following gametes from the females and males:

Females			*Males*		
bw^+	st^+	1/4	bw^+	st^+	1/2
bw^+	st	1/4	bw	st	1/2
bw	st^+	1/4			
bw	st	1/4			

Combining these gametes will give the ratio presented in the tabe of results.

32. The data show that the *ebony* locus is either far away from *B* and *m* (50 map units or more) or that it is on a different chromosome. In fact, *ebony* is on a different chromosome. The data also show linkage of *B* and *m*. Mapping the distance between *B* and *m* would be as follows:

$$(57 + 64)/(226 + 218 + 57 + 64) \times 100$$
$$= 121/565 \times 100 = 21.4 \text{ map units}$$

34. (a) Data presented in this table show an inverse correlation between recombination frequency and live-born children having various trisomies. As the frequency of crossing over decreases, the frequency of trisomy increases. While these data indicate a correlation, other factors such as intrauterine survival are also likely to play a role in determining trisomy live-born frequencies.

(b) If positive interference does spread out crossovers among and within chromosomes, then ensuring that crossovers are distributed among all the chromosomes (and all portions of chromosomes) may reduce nondisjunction and therefore be of selective advantage. This model assumes that the total number of crossovers per oocyte is limiting.

CHAPTER 6

Answers to Now Solve This

6.1 *T C H R O M B A K*

All the genes can be linked together to give a consistent map, and the ends overlap, indicating that the map is circular. The order is reversed in two of the crosses, indicating that the orientation of transfer is reversed.

6.2 Multiplying 0.031×0.012 gives 0.00037 or approximately 0.04 percent. From this information, one would consider no linkage between these two loci. Notice that this frequency is approximately the same as the frequency in the second experiment, where the loci are transformed independently.

6.3 For group A, *d* and *f* are in the same complementation group (gene), while *e* is in a different one. Therefore, $e \times f =$ lysis.

For group B, all three mutations are in the same gene; hence, $b \times i =$ no lysis.

In group C, *j* and *k* are in different complementation groups as are *j* and *l*. It would be impossible to determine whether *l* and *k* are in the same or different complementation group if the *rII* region had more than two cistrons. However, because only two complementation regions exist, and both are not in the same one as *j*, *k* and *l* must both be in the other. Therefore, $k \times l =$ no lysis.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of conjugation, transformation, transduction, and the potential recombination that occurs as a result of these processes.

4. (a) In an $F^+ \times F^-$ cross, the transfer of the F factor produces a recipient bacterium, which is F^+. Any gene may be transferred on an F', and the frequency of transfer is relatively low. Crosses that are Hfr $\times$ F^- produce recombinants at a higher frequency than the $F^+ \times F^-$ cross. The transfer is oriented (nonrandom), and the recipient cell remains F^-.

(b) Bacteria that are F^+ possess the F factor, while those that are F^- lack the F factor. In Hfr cells the F factor is integrated into the bacterial chromosome, and in F' bacteria, the F factor is free of the bacterial chromosome, yet it possesses a piece of the bacterial chromosome.

6. Participating bacteria typically consist of two types, prototrophs and auxotrophs. If a minimal medium is used first, the auxotrophs would be unable to grow and prototrophic recombinants would not be created. If participating bacteria were transferred to a complete medium second, recombinant bacteria would not be distinguishable from donor or recipient bacteria since both prototrophs and auxotrophs grow on a complete medium.

8. The F^+ element can enter the host bacterial chromosome and, upon returning to its independent state, may pick up a piece of a bacterial chromosome. When combined with a bacterium with a complete chromosome, a partial diploid, or merozygote, is formed.

10. During transformation, incoming DNA forms a complex with the host chromosome leading to a double helix that contains one host DNA strand and one incoming DNA strand. Since these two single strands are different in base sequence (otherwise genetic recombination would not occur), the term *heteroduplex* seems appropriate. After one round of replication, the heteroduplex DNA is resolved into two different double-stranded daughter molecules—one with a structure identical to the host's original DNA and one double-stranded mutant DNA that contains the transferred genes.

12. A *plaque* results when bacteria in a "lawn" are infected by a phage and the progeny of the phage destroy (lyse) the bacteria, creating a somewhat clear region called a plaque. *Lysogeny* is a complex process whereby certain temperate phages can enter a bacterial cell and, instead of following a lytic developmental path, integrate their DNA into the bacterial chromosome. In doing so, the bacterial cell becomes lysogenic. The latent, integrated phage chromosome is called a *prophage*.

14. The translation machinery of the infected bacterium provides the necessary materials for protein synthesis.

16. (a) Remembering that 0.1 mL is typically used in the plaque assay, the initial concentration of phage per milliliter is greater than 10^5.

(b) The initial concentration of phage per milliliter is around 140×10^5 or 1.4×10^7.

(c) The initial concentration of phage is less than 10^7. Coupling this information with the calculations in part (b), it would appear that the initial concentration of phage is around 1×10^7 and the failure to obtain plaques in this portion of the experiment is expected and due to sampling error.

18. Groupings into the two complementation groups of the *rII* region are as follows: Group A: *1*, *4*, *5*. Group B: *2*, *3*. Therefore, the result of testing is

$$2 \times 3 = \text{no lysis}$$
$$2 \times 4 = \text{lysis}$$
$$3 \times 4 = \text{lysis}$$

(a) The recombination frequency $= 2(5 \times 10^1)/(2 \times 10^5) = 5 \times 10^{-4}$.

(b) Because mutant *6* complemented mutations *2* and *3*, it is likely to be in the cistron with mutants *1*, *4*, and *5*. A lack of recombinants with mutant *4* indicates that mutant *6* is a deletion that overlaps mutation *4* or it is extremely close to mutant *4*. Recombinants with *1* and *5* indicate that the deletion does not overlap these mutations.

22. (a) The recombination frequency is $2(8 \times 10^2/4 \times 10^7) = 4 \times 10^{-5}$.

(b) The dilution would be 10^{-3} and the plaque number would be 8×10^3.

(c) Mutant *7* might well be a deletion spanning parts of both A and B cistrons.

24. (a) Rifampicin eliminates the donor strain, which is *rif*^{s}.

(b) b_a c_ F

(c) Use a donor strain that was *rif*^{r} but sensitive to another antibiotic (ampicillin, for example). Conduct the interrupted mating experiment as usual on an ampicillin-containing medium, but replate the recombinants on a rifampicin medium to determine which ones are sensitive.

26. If two genes are cotransforming at a relatively high rate, they are said to be "linked." The data indicate that *a* and *d* are linked; *b* and *c* are linked; and since *f* cotransforms with *b*, *b*, *c*, and *f* are likely to be linked. However, if the arrangement is *c b f*, or the reverse, there is a possibility that whereas both *c* and *f* are "linked" to *b*, *c* and *f* may not be strongly linked enough to cotransform. Gene *e* does not cotransform with any gene, so it must be independent of the other linkage groups.

28. (a) Some strains undergo relatively low transfer as a donor strain, whereas others undergo relatively frequent transfer as a donor strain. Within-species transfer is not necessarily more frequent than between-species transfer. The direction of transfer (which are the donor and recipient strains) in some cases influences the frequency of transfer.

(b) *E. chrysanthemi* (–2.4); *E. chrysanthemi* to *E. coli* (–1.7)

(c) Conjugative plasmids can share genes when bacteria are in proximity, and since such plasmids may contain either pathological genes or genes that compromise the use of antibiotics, any harmful variant that develops in one species may be spread to others. Although a particular gene may be harmless in one bacterium, it may confer pathogenicity or drug resistance to a different species.

CHAPTER 7

Answers to Now Solve This

7.1 (a) Something is missing from the male-determining system of sex determination at the level of the genes, gene products, or receptors, etc.

(b) The *SOX9* gene, or its product, is probably involved in male development. Perhaps it is activated by *SRY*.

(c) There is probably some evolutionary relationship between the *SOX9* gene and *SRY*. There is considerable evidence that many other genes and pseudogenes are also homologous to *SRY*.

(d) Normal female sexual development does not require the *SOX9* gene or gene product(s).

7.2 Because of X chromosome inactivation in mammals, scientists would be interested in determining whether the nucleus taken from Rainbow (donor) would continue to show such inactivation. Would the inactivated X chromosome retain the property of inactivation? The white patches of CC are due to an autosomal gene *S* for white spotting that prevents pigment formation in the cell lineages in which it is expressed. Homozygous *SS* cats have more white than heterozygous *Ss* cats, and there is no absolute pattern of patches due to the *S* allele. So, the distribution of white patches would be expected to be different from that in Rainbow. In addition, because X chromosome inactivation is random, CC would have a different patch pattern from that of her genetic mother based on the random X-inactivation basis alone.

Solutions to Problems and Discussion Questions

2. Your essay should include various aspects of sex chromosomes that contain genes responsible for sex determination. Mention should also be made of those organisms in which autosomes play a role in concert with the sex chromosomes.

4. In XX/XO sex determination, sex is determined by the presence or absence of a chromosome, whereas in the XX/XY scheme, the Y chromosome may be sex determining. In some cases, however, the autosomes play a role in sex determination. The *Protenor* form of sex determination involves the XX/XO condition, while the *Lygaeus* mode involves the XX/XY condition.

6. Mammals possess a system of X chromosome inactivation whereby one of the two X chromosomes in females becomes a chromatin body or Barr body. If one of the two X chromosomes is randomly inactivated, the dosage of genetic information is more or less equivalent in males (XY) and females (XX).

8. In humans, individuals with the 47,XXY complement are males, whereas those with 45,XO are females. Thus, the Y chromosome—specifically, the sex-determining region (*SRY*)—was deduced as male determining in humans. *SRY* encodes a product that causes the undifferentiated gonadal tissue to form testes. Thus, individuals who are XX but with the *SRY* gene translocated to one X chromosome develop as males and XY individuals with the *SRY* gene deleted or nonfunctional are phenotypically female. By contrast, XO *Drosophila* are males and XXY individuals are female. Studies on the offspring of triploid females led to the realization that it is the balance between the number of X chromosomes and the number of haploid sets of autosomes that determines sex in *Drosophila*.

10. (a) female $X^{rw}\,Y \times$ male $X^{+}X^{+}$

F_1:	females:	$X^{+}\,Y$ (normal)
	males:	$X^{rw}\,X^{+}$ (normal)
F_2:	females:	$X^{+}\,Y$ (normal)
		$X^{rw}\;Y$ (reduced wing)
	males:	$X^{rw}\;X^{+}$ (normal)
		$X^{+}\;X^{+}$ (normal)

(b) female $X^{rw}\,X^{rw} \times$ male $X^{+}\;Y$

F_1:	females:	$X^{rw}\;X^{+}$ (normal)
	males:	$X^{rw}\;Y$ (reduced wing)
F_2:	females:	$X^{rw}\;X^{+}$ (normal)
		$X^{rw}\;X^{rw}$ (reduced wing)
	males:	$X^{+}\,Y$ (normal)
		$X^{rw}\;Y$ (reduced wing)

12. Because attached-X chromosomes have a mother-to-daughter inheritance and the father's X is transferred to the son, one would see daughters with the white-eye phenotype and normal wings and sons with the miniature wing phenotype and red eyes. In addition, there would be rare ($<3\%$) metafemales (attached-X + X) with wild-type eye color and normal wings. YY zygotes fail to develop into larvae.

14. Because synapsis of chromosomes in meiotic tissue is often accompanied by crossing over, it would be detrimental to sex-determining mechanisms to have sex-determining loci on the Y chromosome transferred, through crossing over, to the X chromosome.

16.

Klinefelter syndrome (XXY)	= 1
Turner syndrome (XO)	= 0
47,XYY	= 0
47,XXX	= 2
48,XXXX	= 3

18. Unless other markers, cytological or molecular, are available, one cannot test the Lyon hypothesis with homozygous X-linked genes.

20. Dosage compensation and the formation of Barr bodies occur only when there are two or more X chromosomes. Males normally have only one X chromosome; therefore, such mosaicism cannot occur.

22. Because of the homology between the *red* and *green* genes, there exists the possibility for an irregular synapsis, which, following after crossing over, would give a chromosome with only one (*green*) of the duplicated genes.

24. Since all haploids are male and half of the eggs are unfertilized, 50 percent of the offspring would be male at the start; adding the X_a/X_a types gives 25 percent more male; the remainder X_a/X_b would be female. Overall, 75 percent of the offspring would be male, and 25 percent would be female.

26. The presence of the Y chromosome provides a factor (or factors) that leads to the initial specification of maleness and the formation of testes. Subsequent expression of secondary sex characteristics must be dependent on the interaction of the normal X-linked *Tfm* allele with testosterone. Without such interaction, differentiation takes the female path. To test the dominant nature of the *Tfm* allele, one could generate an XXY male that is heterozygous for the *Tfm* allele. Should the same testicular feminization phenotype occur, the dominant nature of the *Tfm* allele would be supported. If the *Tfm* phenotype was eliminated in the heterozygote, one would have support for the model that the normal *Tfm* allele is needed to interact with testosterone.

28. As the climate warms, more genotypically *ZZ* individuals will develop as females rather than as males. This will tend to skew the population toward a male-to-female sex ratio of less than one. Further, *ZZ* females can mate with *ZZ* males, producing only *ZZ* offspring. Eventually, the population might consist only of genotypically *ZZ* individuals and might rely upon temperature-dependent sex determination.

CHAPTER 8

Answers to Now Solve This

8.1 If the father had hemophilia, it is likely that the Turner syndrome individual inherited the X chromosome from the father and no sex chromosome from the mother. If nondisjunction occurred in the mother, during either meiosis I or meiosis II, an egg with no X chromosome can be the result.

8.2 The sterility of interspecific hybrids is often caused by a high proportion of univalents in meiosis I. As such, viable gametes are rare and the likelihood of two such gametes "meeting" is remote. Even if partial homology of chromosomes allows some pairing, sterility is usually the rule. The horticulturist may attempt to reverse the sterility by treating the sterile hybrid with colchicine. Such a treatment, if successful, may double the chromosome number, and each chromosome would then have a homolog with which to pair during meiosis.

8.3 The rare double crossovers *within the boundaries of a paracentric or pericentric inversion* produce only minor departures from the standard chromosomal arrangement as long as the crossovers involve the same two chromatids. With two-strand double crossovers, the second crossover negates the first. However, three-strand and four-strand double crossovers have consequences that lead to anaphase bridges as well as a high degree of genetically unbalanced gametes.

Solutions to Problems and Discussion Questions

2. Your essay can draw from many examples discussed in the text including deletions, duplications, inversions, translocations, and copy number variations.

4. haploid = 9, tetraploid = 36,
trisomic = 19, monosomic = 17

6. Karyotype analysis of spontaneously aborted fetuses has shown that a significant percentage of abortuses are trisomic and any of the chromosomes can be involved. Other forms of aneuploidy (monosomy, nullisomy) are less represented.

8. American cultivated cotton has 26 pairs of chromosomes: 13 large and 13 small. Old World cotton has 13 pairs of large chromosomes, and American wild cotton has 13 pairs of small chromosomes. It is likely that an interspecific hybridization occurred followed by chromosome doubling. These events probably produced a fertile amphidiploid (allotetraploid). Experiments have successfully been conducted to reconstruct the origin of American cultivated cotton.

10. While there is the appearance that crossing over is suppressed in inversion "heterozygotes," the phenomenon extends from the fact that the crossover chromatids end up being abnormal in genetic content. As such, they fail to produce viable (or competitive) gametes, or they lead to zygotic or embryonic death.

12. Modern globin genes resulted from a duplication event in an ancestral gene about 500 million years ago. Mutations occurred over time and a chromosomal aberration separated the duplicated genes, leaving the eventual α gene cluster on chromosome 16 and the eventual β gene cluster on chromosome 11.

14. By having the genes in an inversion, crossover chromatids are not recovered and therefore are not passed on to future generations. Translocations offer an opportunity for new gene combinations by associations of genes from nonhomologous chromosomes. Under certain conditions, such new combinations may be of selective advantage. Meiotic conditions have evolved so that segregation of translocated chromosomes yields a relatively uniform set of gametes.

16. Given the basic chromosome set of 9 unique chromosomes (a haploid complement), other forms with the "*n* multiples" are forms of autopolyploidy. Individual organisms with 27 chromosomes (3*n*) are more likely to be sterile because there are trivalents at meiosis I, which cause a relatively high number of unbalanced gametes to be formed.

18. The cross would be as follows:

WWWW × *wwww*
(assuming that chromosomes pair
as bivalents at meiosis)

F$_1$: *WWww*

F$_2$: 1 *WW* 4 *Ww* 1 *ww*

1*WW*
4*WW*
1*ww*

Fill in Punnett square
35 *W* phenotypes and 1 *w* phenotype.

20. While a number of mechanisms for *bobbed* reversion have been documented, one based on meiotic recombination occurs through "unequal crossing over." When redundant chromosomal regions synapse, homologs can misalign. If crossing over occurs in the misaligned segments, one chromatid can gain chromosomal material at the expense of the other chromatid. As a chromosome gains rRNA genes, it harbors a selective advantage and produces flies that outcompete those with nonreverted *bobbed* mutations. Eventually, a stock that originally contained *bobbed Drosophila* contains what appear to be wild-type flies over time.

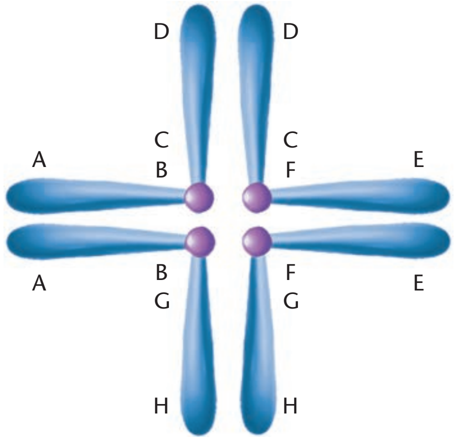
22. (a) In all probability, crossing over in the inversion loop of an inversion (in the heterozygous state) had produced defective, unbalanced chromatids, thus leading to stillbirths and/or malformed children.

(b) It is probable that a significant proportion (perhaps 50 percent) of the children of the man will be similarly influenced by the inversion.

(c) Since the karyotypic abnormality is observable, it may be possible to detect some of the abnormal chromosomes of the fetus by amniocentesis or CVS. However, depending on the type of inversion and the ability to detect minor changes in banding patterns, not all abnormal chromosomes may be detected.

24. (a) Reciprocal translocation

(b)



(c) Notice that all chromosomal segments are present and there is no apparent loss of chromosomal material. However, if the breakpoints for the translocation occurred within genes, then an abnormal phenotype may be the result. In addition, a gene's function is sometimes influenced by its position (its neighbors, in other words). If such "position effects" occur, then a different phenotype may result.

26. The symbolism t(14;21) indicates that a translocation (t) has occurred between chromosomes 14 and 21. Generally, a Down syndrome individual with a t(14;21) karyotype has 46 chromosomes.

28. Loss through nondisjunction of one of the sister chromatids during mitosis could have occurred, thereby producing two daughter cells, one 46,XX and the other 45,XO. The earlier the event occurred during embryogenesis, the greater the percentage of 45,XO tissue is expected. The distribution of the 45,XO tissue would be dependent on the eventual fate of cells where the nondisjunctional event occurred.

30. This female will produce meiotic products of the following types:

normal: 15 + 21
translocated: 15/21
translocated plus 21: 15/21 + 21
deficient:15 only

Note: The 15/21 + 15 gamete is not formed, because it would require separation of primarily homologous chromosomes at anaphase I.

Fertilization with a normal 15 + 21 sperm cell will produce the following offspring:

normal: 46 chromosomes
translocation carrier: 45 chromosomes 15/21 + 15 + 21
trisomy 21:46 chromosomes 15/21 + 21 + 21
monosomic: 45 chromosomes 15 + 15 + 21, lethal

CHAPTER 9

Answers to Now Solve This

9.1 The mt^+ strain is the donor of the *cp*DNA since the inheritance of resistance or sensitivity is dependent on the status of the mt^+ gene. In this organism, chloroplasts obtain their characteristics from the mt^+ strain, while mitochondria obtain their characteristics from the mt^- strain.

9.2 (a) neutral
(b) segregational (nuclear mutations)
(c) suppressive

9.3 In many cases, molecular components of mitochondria are recruited from the cytoplasm, having been synthesized from nuclear genes.

9.4 From an organismic standpoint, individuals with the most severe mitochondrial defects tend to be less reproductively successful. At the cellular level, mitochondria with mutations in protein-coding genes tend to be selected against. Such purifying selection tends to favor nonmutant mitochondria.

Solutions to Problems and Discussion Questions

2. Your essay should be based on the fact that in organelle heredity, an organelle is responsible for the inheritance pattern, whereas in a maternal effect, organelles are not involved. Maternal effects persist for one generation only and are solely dependent on the genotype of the mother.

4. Because the ovule source furnishes the cytoplasm, and thus the chloroplasts to the embryo, the offspring will have the same phenotype as the plant providing the ovule.
(a) green
(b) white
(c) white, variegated, or green
(d) green

6. If the two are crossed as stated in the problem, then the diploid zygote would contain mitochondria with normal mtDNA and would be heterozygous for the nuclear *segregational petite* gene. This mutant allele would be "covered" by the normal nuclear allele from the neutral strain, resulting in a normal phenotype. During meiosis, normal mitochondria would be distributed to all ascospores, whereas the nuclear genes would segregate in a 1 : 1 ratio. Fifty percent of the haploid ascospores would be normal, and 50 percent would express the *segregational* allele and be petite.

8. The fact that all the offspring (F_1) showed a dextral coiling pattern indicates that one of the parents (maternal parent) contains at least one copy of the *D* allele. Taking these offspring and seeing that their progeny (call these F_2) occur in a 1 : 1 ratio indicates that half of the offspring (F_1) are *dd*. In order to have these results, one of the original parents must have been *Dd*, while the other must have been *dd*.

10. Since there is no evidence for segregation patterns typical of chromosomal genes and Mendelian traits, some form of extranuclear inheritance seems possible. If the *lethargic* gene is dominant, then a maternal effect may be involved. In that case, some of the F_2 progeny would be hyperactive because maternal effects are temporary, affecting only the immediate progeny. If caused by a mitochondrial defect, then the condition would persist in all offspring of lethargic mothers, through more than one generation.

12. The endosymbiotic theory states that mitochondria and chloroplasts arose independently around 2 billion years ago from free-living protobacteria. These bacteria brought the capacity for aerobic respiration and photosynthesis to primitive eukaryotic cells. Because such organelles have bacterial origins, a deeper understanding of extranuclear DNA is possible.

14. MRT is a highly promising technique for eliminating the risk of bearing children with diseases due to defective genes in the maternal mitochondrial DNA. In theory, using MRT results in all mitochondria being derived from a healthy egg-donor and nuclear genes being contributed by the parents. Currently, however, the technology is not yet able to prevent the co-transfer of some defective mitochondria with the maternal nucleus. It is not yet known whether such low-level transfers would have health consequences for the child.

16. (a) The presence of bcd^-/bcd^- males can be explained by the maternal effect: mothers were bcd^+/bcd^- and were able to provide the bicoid protein to the egg.
(b) The cross: female bcd^+/bcd^- × male bcd^-/bcd^- will produce an F_1 with normal embryogenesis because of the maternal effect. In the F_2, any cross having bcd^+/bcd^- mothers will have phenotypically normal embryos. Offspring from any cross involving homozygous bcd^-/bcd^- mothers will fail to develop anterior structures, due to the absence of the bicoid protein in the egg.

18. (a) Since these disorders involve pools of defective mitochondrial DNA in virtually all cells of the body and a certain percentage of healthy mitochondria is necessary for normal function, developing a cure is difficult. Defective mitochondria would need to be corrected, which appears to be technologically impossible at this time. An alternative approach might be suppression of the mutant mitochondrial replication and favoring the replication of normal mitochondria.
(b) Mitochondria can be successfully microinjected into mammalian oocytes, where they reproduce and behave as typical, ATP-producing organelles. Through this method, it may be possible to alter the heteroplasmic ratio by microinjection, thereby reducing the likelihood of a child being born with a mitochondrial disease.

20. (a) The two sources of heteroplasmy are likely to be new mutations and/or cytoplasmic inheritance from the mother.
(b) The most likely genetic condition within a given mitochondrion would be some sort of mutator gene such as a faulty polymerase.

22. Evidence supporting the so-called Mother's curse would include decreased male fitness (measured by viability, aging, fecundity, etc.) versus females with identical genetic backgrounds. Data comparing the effect of different mtDNAs on variations in nuclear gene expression in males versus females would be of interest. Genes to observe would include those related to sex-specific fitness.

CHAPTER 10

Answers to Now Solve This

10.1 In an *in vitro* experiment like that of Avery et al., in theory, the general design would be appropriate: in that some substance, if labeled, would show up in the progeny of transformed bacteria. However, since the amount of transforming DNA is extremely small compared with the genomic DNA of the recipient bacterium and its progeny, it would be technically difficult to assay for the labeled nucleic acid. In addition, it would be necessary to know that the small stretch of DNA that caused the genetic transformation was actually labeled.

10.2 guanine = 17.5%; adenine and thymine both each = 32.5%

10.3 Since uracil is present rather than thymine, the genetic material is RNA. Assuming that the value of 1.13 is statistically different from 1.00, purines outnumber pyrimidines; therefore, the nucleic acid is single stranded. Overall, one can conclude that rubella is a single-stranded RNA virus.

Solutions to Problems and Discussion Questions

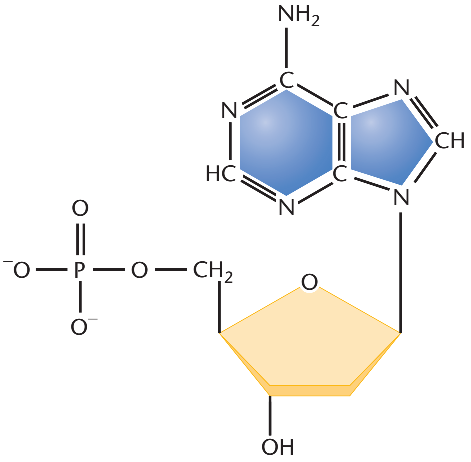
2. Your essay should include a description of structural aspects including sugar and base content comparisons. In addition, you should mention complementation aspects, strandedness, flexibility, and conformation.

4. Griffith was the first to observe (and name) the phenomenon of transformation, using an *in vivo* system (laboratory mice). He observed that a bacterial mixture containing heat-killed cells of a virulent strain of *Diplococcus pneumoniae* and live cells of an avirulent strain killed the injected mice and led to recovery of live cells of the virulent strain. He concluded that the heat-killed virulent bacteria transformed the avirulent strain into a virulent strain. He did not identify the "transforming principle." By contrast, Avery and coworkers, using *in vitro* techniques, systematically searched for the transforming principle originating from the heat-killed pathogenic strain and determined it to be DNA.

6. Nucleic acids contain large amounts of phosphorus and no sulfur, whereas proteins contain sulfur and no phosphorus. Therefore, the radioisotopes ^{32}P and ^{35}S will selectively label nucleic acids and proteins, respectively. The Hershey and Chase experiment was based on the premise that the substance injected into the bacterium is the substance responsible for producing the progeny phage and therefore must be the hereditary material. The experiment demonstrated that most of the ^{32}P-labeled material (DNA) was injected, while the ^{35}S-labeled phage ghosts (protein coats) remained outside the bacterium. Therefore, the nucleic acid must be the genetic material.

8. The early evidence that DNA was the genetic material would be considered indirect. The fact that DNA was located only in subcellular structures where genetic functions occur (the nucleus, chloroplasts and mitochondria) favored DNA over proteins as the genetic information. Further indirect evidence came from the observation that DNA content and ploidy in various cell types (sperm and somatic cells) were related and that the *action* and *absorption* spectra of ultraviolet light were correlated. Direct evidence for DNA being the genetic material comes from a variety of observations including gene transfer, which has been facilitated by recombinant DNA techniques.

10. The structure of deoxyadenylic acid is given below and in the text. Linkages among the three components require the removal of water (H_2O).



12.		
12.	guanine:	2-amino-6-oxypurine
	cytosine:	2-oxy-4-aminopyrimidine
	thymine:	2,4-dioxy-5-methylpyrimidine
	uracil:	2,4-dioxypyrimidine

14. The following are characteristics of the Watson-Crick double-helix model for DNA: There are two polynucleotide chains, each formed by phosphodiester linkages between the five-carbon sugars and the phosphates. Bases are stacked 0.34 nm (3.4 angstroms) apart and in a plectonic, antiparallel manner. There is one complete turn for each 3.4 nm, which constitutes 10 bases per turn. Hydrogen bonds hold the two polynucleotide chains together. There are two hydrogen bonds forming the A-T pair and three forming the G-C pair. The double helix exists as a twisted structure, approximately 20 angstroms in diameter, with a topography of major and minor grooves. The hydrophobic bases are located in the center of the molecule; the hydrophilic phosphodiester backbone is on the outside.

16. Because in double-stranded DNA, A-T and G-C (within limits of experimental error), the data presented would have indicated a lack of pairing of these bases in favor of a single-stranded structure or some other non-hydrogen-bonded structure. Alternatively, from the data it would appear that A = C and T = G, which would negate the chance for typical hydrogen bonding since opposite charge relationships do not exist. Therefore, it is quite unlikely that a tight helical structure would form at all. In conclusion, Watson and Crick might have concluded that hydrogen bonding is not a significant factor in maintaining a double-stranded structure.

18. Three main differences between RNA and DNA are the following:

(1) Uracil in RNA replaces thymine in DNA.

(2) Ribose in RNA replaces deoxyribose in DNA.

(3) RNA often occurs as both single-stranded and partially double-stranded forms, whereas DNA most often occurs in a double-stranded form.

20. Nucleic acids absorb UV light maximally at wavelengths of 254–260 nm. Using this phenomenon, one can often determine the presence and concentration of nucleic acids in a mixture since proteins absorb UV light maximally at 280≈nm. UV absorption is greater in single-stranded molecules as compared with double-stranded structures (hyperchromic shift); therefore, one can easily determine, by applying denaturing conditions, whether a nucleic acid is in the single- or double-stranded form. In addition, A-T—rich DNA denatures more readily than G-C—rich DNA; therefore, one can estimate base content by denaturation kinetics.

22. A *hyperchromic effect* is the increased absorption of UV light as double-stranded DNA (or RNA for that matter) is converted to single-stranded DNA. As illustrated in the text, the change in absorption is quite significant, with a structure of higher G-C content *melting* at a higher temperature than an A-T—rich nucleic acid. If one monitors the UV absorption with a spectrophotometer during the melting process, the hyperchromic shift can be observed. The T_m is the point on the profile (temperature) at which half (50 percent) of the sample is denatured.

24. The reassociation of separate complementary strands of a nucleic acid, either DNA or RNA, is based on hydrogen bonds forming between A-T (or U) and G-C.

26. (1) As shown, the extra phosphate is not normally expected.

(2) In the adenine ring, a nitrogen is at position 8 rather than position 9 and a carbon is at position 9 rather than position 8.

(3) The bond from the $C - 1'$ to the sugar should form with the N at position 9 (N-9) of the adenine.

(4) There should be a double bond between C-4 and C-5 of adenine.

(5) The dinucleotide is a "deoxy" form; therefore, each $C - 2'$ should not have a hydroxyl group. Notice the hydroxyl group at $C - 2'$ on the sugar of the adenylic acid.

(6) At the C-5 position on the thymine residue, there should be a methyl group.

(7) There are too many bonds between the N-3—C-2 of thymine.

(8) There are too few bonds (should be a double bond) between the C-5 and C-6 of thymine.

(9) The extra hydroxyl group on $C - 5'$ of the sugar of the thymidylic acid should not be there (the $C - 5'$ hydroxyl group is involved in the bond with the phosphate group).

28. Since cytosine pairs with guanine and uracil pairs with adenine, the result would be a base substitution of G:C to A:T after two rounds of replication.

30. Fluorescence *in situ* hybridization employs fluorescently labeled DNA that hybridizes to metaphase chromosomes and interphase nuclei. A FISH survey is considered interpretable if hybridization is consistent in 70 percent or more of cells examined. Results are available in one to two days after the sample is tested. Because of the relatively high likelihood of aneuploidy for chromosomes 13, 18, 21, X, and Y, they are routine candidates for analysis.

31. (a—c) Without knowing the exact bonding characteristics of hypoxanthine or xanthine, it may be difficult to predict the likelihood of each pairing type. It is likely that both are of the same class (purine or pyrimidine) because the names of the molecules indicate a similarity. In addition, the diameter of the structure is constant, which, under the model to follow, would be expected. In fact, hypoxanthine and xanthine are both purines.

Because there are equal amounts of A, T, and H, one could suggest that they are hydrogen bonded to one another; the same may be said for C, G, and X. Given the molar equivalence of erythrose and phosphate, an alternating sugar-phosphate-sugar backbone, as in "earth-type" DNA, would be acceptable. A model of a triple helix would be acceptable, since the diameter is constant. Given the chemical similarities to "earth-type" DNA, it is probable that the unique creature's DNA follows the same structural plan.

34. The 3′-OH group is critical for both DNA and RNA because it is involved in the phosphodiester bonds that link together deoxynucleotides or nucleotides into the long polymers that function as informational molecules. A nucleic acid with a 3′-H group would be unable to form phosphodiester bonds and thus would be unable to form polymers.

CHAPTER 11

Answers to Now Solve This

11.1 After one round of replication in the ^{14}N medium, the conservative scheme can be ruled out. After one round of replication in ^{14}N under a dispersive model, the DNA would be of intermediate density, just as it is in the semiconservative model. However, after the next round of replication in ^{14}N medium, the density of the DNA is between the intermediate and "light" densities, thus ruling out the dispersive model.

11.2 If the DNA contained parallel strands in the double helix and the polymerase would be able to accommodate such parallel strands, there would be continuous synthesis and no Okazaki fragments. Several other possibilities exist. If the DNA were replicated as single strands, the synthesis could begin at the free ends and there would be no need for Okazaki fragments.

Solutions to Problems and Discussion Questions

2. Your essay should describe replication as the process of making daughter nucleic acids from existing ones. Synthesis refers to the precise series of steps, components, and reactions that allow such replication to occur.

4. The Meselson and Stahl experiment has the following components. By labeling the pool of nitrogenous bases of the DNA of *E. coli* with heavy isotope ^{15}N, it would be possible to "follow" the "old" DNA. Cells were grown for many generations in medium containing ^{15}N and then transferred to ^{14}N medium so that "new" DNA could be detected. A comparison of the density of DNA samples at various times in the experiment (initial ^{15}N culture and subsequent cultures grown in the ^{14}N medium) showed that after one round of replication in the ^{14}N medium, the DNA was half as dense (intermediate) as the DNA from bacteria grown only in the ^{15}N medium. In a sample taken after two rounds of replication in the ^{14}N medium, half of the DNA was of the intermediate density and the other half was as dense as DNA containing only ^{14}N DNA.

6. The *in vitro* replication requires a DNA template, a divalent cation (Mg^{2+}), and all four of the deoxyribonucleoside triphosphates: dATP, dCTP, dTTP, and dGTP. The lower-case "d" refers to the deoxyribose sugar.

8. The indirect approach described in the text was analysis of *base composition*, which indicated that the composition of the product met the expected composition within experimental error.

10. An exposed 3′-OH group is necessary for the attachment of the next nucleotide. The 3′-OH group is eventually removed in the form of water, and a covalent bond is formed to the 5′-phosphate of the added nucleotide.

12. All three enzymes share several common properties. None can *initiate* DNA synthesis on a template, but all can *elongate* an existing DNA strand assuming there is a "primer" strand with a free 3′-OH annealed to a longer template strand. Polymerization of nucleotides occurs in the 5′ to 3′ direction where each 5′ phosphate is added to the 3′ end of the growing polynucleotide.

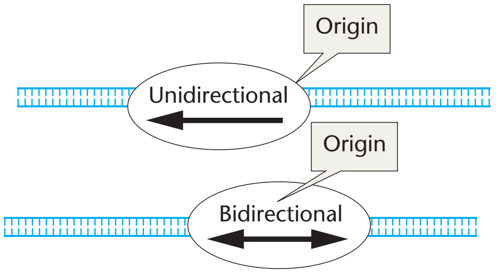
All three enzymes are large, complex proteins with a molecular weight in excess of 100,000 daltons, and each has 3′ to 5′ exonuclease activity. Refer to the text.

DNA polymerase I:
5′ to 3′exonuclease activity
Present in large amounts
Relatively stable
Removal of RNA primer

DNA polymerase II:
Possibly involved in repair function

DNA polymerase III:
Essential for replication
Complex molecule

14. Given a stretch of double-stranded DNA, one could initiate synthesis at a given point and either replicate strands in one direction only (unidirectional) or in both directions (bidirectional) as shown below. Notice that in the text the synthesis of complementary strands occurs in a *continuous* 5′ to 3′ mode on the leading-strand template in the direction of the replication fork, resulting in a single, long product strand, and in a *discontinuous* 5′ to 3′ mode on the lagging strand resulting in shorter product strands called Okazaki fragments.



16. (a) *Okazaki fragments* are relatively short (1000 to 2000 bases in bacteria) DNA fragments that are synthesized in a discontinuous fashion on the lagging-strand templates during DNA replication. Such fragments appear to be necessary because template DNA is not available for 5′ to 3′ synthesis until some degree of continuous DNA synthesis occurs on the leading-strand template in the direction of the replication fork. The isolation of such fragments provides support for the scheme of replication shown in the text.

(b) DNA *ligase* is required to form phosphodiester linkages in nicks, which are generated when DNA polymerase I removes RNA primer and meets newly synthesized DNA ahead of it. The discontinuous DNA strands are ligated together into a single continuous strand.

(c) *Primer* RNA is formed by RNA primase to serve as an initiation point for the production of DNA strands on a DNA template. None of the DNA polymerases are capable of initiating synthesis without a free 3′ hydroxyl group. The primer RNA provides that group and thus can be used by DNA polymerase III.

18. Eukaryotic DNA is replicated in a manner that is very similar to that of *E. coli*. Synthesis is bidirectional and continuous on one strand and discontinuous on the other, and the requirements of synthesis (four deoxyribonucleoside triphosphates, divalent cation, template, and primer) are the same. Okazaki fragments of eukaryotes are about one-tenth the size of those in bacteria.

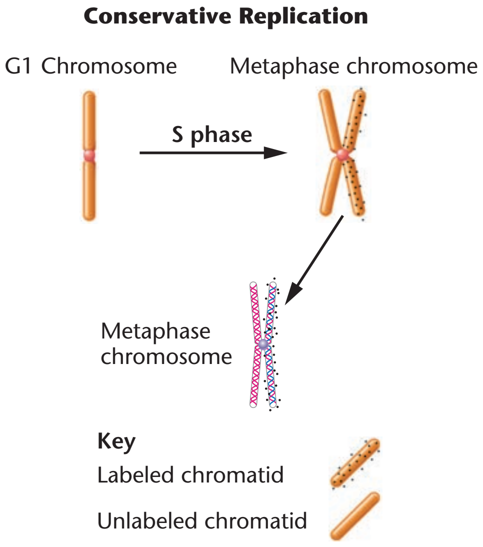
Because there is a much greater amount of DNA to be replicated and DNA replication is slower, there are multiple initiation sites for replication in eukaryotes (and increased DNA polymerase per cell) in contrast to the single replication origin in bacteria. Replication occurs at different sites during different intervals of the S phase.

20. (a) No repair from DNA polymerase I and/or DNA polymerase III
(b) No DNA ligase activity
(c) No primase activity
(d) Only DNA polymerase I activity
(e) No DNA gyrase activity

22. The end-replication problem refers to the difficulties posed in replicating the ends of linear eukaryotic chromosomes. Once primers are removed from the 5′ ends, a gap remains, which cannot be filled. This shortens the chromosome with each round of replication, potentially leading to deletion of gene-coding regions. The action of telomerase lengthens the telomere, which is then made double stranded (except for the end) through conventional DNA synthesis.

24. Telomerase activity is present in germ-line tissue to maintain telomere length from one generation to the next. It is also necessary in stem cells and other proliferating tissues. In other words, telomeres cannot shorten indefinitely without eventually eroding genetic information.

26. If replication is conservative, radioactive label would be distributed on only one side (chromatid) of the metaphase chromosome in the autoradiographs of metaphase I, as shown below.



28. (a) 5′ACCUAAGU-3′
(b) U

30. (a) DNA, since one of the nitrogenous bases is T; also, notice the lack of an OH group at the 2′ carbon.
(b) 3′.
(c) Since spleen diesterase cuts between the 5′ carbon and the phosphate, the original 5′ phosphate is transferred to the 3′ carbon of the 5′ neighbor. Therefore, deoxyadenosine would obtain the phosphate at its 3′position.

CHAPTER 12

Answers to Now Solve This

12.1 By having a circular chromosome, there are no free ends to present the problem of linear chromosomes, namely, complete replication of terminal sequences.

12.2 Since polytene chromosomes are formed by multiple rounds of DNA replication without strand separation, you would expect grains along the entire length of each polytene chromosome.

12.3 Volume of the nucleus $= 4/3\pi r^3$

$$= 4/3 \times 3.14 \times (5 \times 10^3\ \text{nm})^3$$
$$= 5.23 \times 10^{11}\ \text{nm}^3$$

Volume of the chromosome $= \pi r^2 \times$ length

$$= 3.14 \times 5.5\ \text{nm} \times 5.5\ \text{nm} \times (2 \times 10^9\ \text{nm})$$
$$= 1.9 \times 10^{11}\ \text{nm}^3$$

Therefore, the percentage of the volume of the nucleus occupied by the chromatin is

$$= (1.9 \times 10^{11}\ \text{nm}^3/5.23 \times 10^{11}\ \text{nm}^3) \times 100$$
$$= \text{about } 36.3\%$$

Solutions to Problems and Discussion Questions

2. Your essay should include descriptions of overall chromosomal configurations, such as linearity or circularity, strandedness, as well as association with chromosomal proteins. In addition, it should describe higher-level structures such as condensation in the case of eukaryotic chromosomes.

4. Polytene chromosomes are formed from numerous DNA replications, pairing of homologs, and absence of strand separation or cytoplasmic division. Each chromosome contains about 1000–5000 DNA strands in parallel register. They are found in specific tissues, such as salivary glands, of many dipterans such as *Drosophila*. They appear as comparatively long, wide fibers with sharp light and dark sections (bands) along their length. Such bands (chromomeres) are useful in chromosome identification, etc.

6. Lampbrush chromosomes are typically present in vertebrate oocytes and in spermatocytes of some insects. They are found during late prophase I and are active uncoiled versions of condensed meiotic chromosomes.

8. Digestion of chromatin with endonucleases, such as micrococcal nuclease, gives DNA fragments of approximately 200 base pairs or multiples of such segments. Digestion for longer times revealed shortened DNA fragments (147 bp) and suggested the presence of linker DNA. Regularly spaced beadlike structures (nucleosomes) were identified by electron microscopy. X-ray diffraction data indicated a regular spacing of DNA in chromatin.

10. As chromosome condensation occurs, a 300-nm fiber is formed. It appears to be composed of 5 or 6 nucleosomes coiled together. Such a structure is called a solenoid. These fibers form a series of loops that further condense into the chromatin fiber, which are then coiled into chromosome arms making up each chromatid.

12. (a) Since there are 200 base pairs per nucleosome (as defined in this problem) and 10^9 base pairs, there would be 5×10^6 nucleosomes.
(b) Since there are 5×10^6 nucleosomes and nine histones (including H1) per nucleosome, there must be $9(5 \times 10^6)$ histone molecules: 4.5×10^7.
(c) Since there are 10^9 base pairs present and each base pair is 0.34 nm, the overall length of the DNA is 3.4×10^8 nm. Dividing this value by the packing ratio (50) gives 6.8×10^6 nm or 6.8 mm.

14. One base pair occupies 0.34 nm; therefore, the equation would be as follows:

$$52\mu\text{m}/(0.34\ \text{nm/bp}) \times 1000\ \text{nm}/\mu\text{m}$$
$$= 152{,}941\ \text{base pairs}$$

16. Whereas SINEs are small (usually less than 500 bp) and present in over one million copies, LINEs are quite large (usually 6 kb in length) and less prevalent. LINEs are often referred to as retrotransposons because their mechanism of transposition (transcription to RNA which is copied back to DNA by reverse transcriptase) resembles that used by retroviruses.

18. In many cases, viruses specifically methylate the genetic apparatus associated with the immune response, thus dampening that response and enhancing viral infectivity.

20. The given data support the general observation that heterochromatic genes are less active than euchromatic genes, and, more specifically, the possibility that heterochromatin may contain genes that are repressed.

22. DNA replicates in a *semiconservative* fashion, with each daughter DNA double helix containing one new and one original single strand. Nucleosomes follow a *dispersive* pattern, with each daughter chromatid containing a mixture of new and original nucleosomes. One could test the distribution of nucleosomes by conducting an autoradiographic experiment similar to Taylor-Woods-Hughes, but instead of labeling the DNA with ^{3}H-thymidine, one would label all or some of the histones H2A, H2B, H3, and H4 in nucleosomes.

24. Bacteriophage λ is composed of a double-stranded, linear DNA molecule of about 48,000 base pairs. It is capable of forming a closed, double-stranded circular molecule because of a 12-base-pair, single-stranded, complementary "overhanging" sequence at the 5′ end of each single strand.

26. Since both genes mentioned in the problem are near the telomeric heterochromatin, they may be subject to silencing due to the position effect. It is also possible that erosion of the end of the chromosome is related to each disease. Examination of the gene by *in situ* hybridization and molecular cloning indicates that thalassemia involves a terminal deletion in the distal portion of 16p. To learn more about such conditions, visit https://www.omim.org/ and follow the OMIM link.

CHAPTER 13

Answers to Now Solve This

13.1 (a) The way to determine the fraction of each triplet that will occur with a random incorporation system is to determine the likelihood that each base will occur in each position of the codon (first, second, third) and then multiply the individual probabilities (fractions) for a final probability (fraction).

GGG = $3/4 \times 3/4 \times 3/4 = 27/64$
GGC = $3/4 \times 3/4 \times 1/4 = 9/64$
GCG = $3/4 \times 1/4 \times 3/4 = 9/64$
CGG = $1/4 \times 3/4 \times 3/4 = 9/64$

CCG = $1/4 \times 1/4 \times 3/4 = 3/64$
CGC = $1/4 \times 3/4 \times 1/4 = 3/64$
GCC = $3/4 \times 1/4 \times 1/4 = 3/64$
CCC = $1/4 \times 1/4 \times 1/4 = 1/64$

(b) Glycine:
GGG and one G_2C (adds up to 36/64)
Alanine:
one G_2C and one C_2G (adds up to 12/64)
Arginine:
one G_2C and one C_2G (adds up to 12/64)
Proline:
one C_2G and CCC (adds up to 4/64)

(c) With the wobble hypothesis, variation can occur in the third position of each codon.

Glycine: GGG, GGC
Alanine: CGG, CGC or GCC, GCG
Arginine: GCG, GCC, or CGC, CGG
Proline: CCC, CCG

13. 2 Assume that you have introduced a copolymer (ACACACAC . . .) to a cell-free protein-synthesizing system. There are two possibilities for establishing the reading frames: ACA, if one starts at the first base; and CAC, if one starts at the second base. These would code for two different amino acids (ACA = threonine; CAC = histidine) and would produce repeating polypeptides that would alternate Thr-His-Thr-His . . . *or* His-Thr-His-Thr . . .

Because of a triplet code, a trinucleotide sequence will, once initiated, remain in the same reading frame and produce the same code all along the sequence regardless of the initiation site.

Given the sequence CUACUACUACUA, notice the different reading frames producing three different sequences, each containing a single, repeated amino acid.

Codons:	CUA	CUA	CUA	CUA. . .
Amino acids:	Leu	Leu	Leu	Leu. . .
	UAC	UAC	UAC	UAC. . .
	Tyr	Tyr	Tyr	Tyr. . .
	ACU	ACU	ACU	ACU. . .
	Thr	Thr	Thr	Thr. . .

If a tetranucleotide is used, such as ACGUACGUACGU . . .

Codons:	ACG	UAC	GUA	CGU	ACG
Amino acids:	Thr	Tyr	Val	Arg	Thr
	CGU	ACG	UAC	GUA	CGU
	Arg	Thr	Tyr	Val	Arg
	GUA	CGU	ACG	UAC	GUA
	Val	Arg	Thr	Tyr	Val
	UAC	GUA	CGU	ACG	UAC
	Tyr	Val	Arg	Thr	Tyr

Notice that the sequences are the same except that the starting amino acid changes.

13.3 Apply complementary bases, substituting U for T:

(a)

Sequence 1: 5′-AUGGCAAAAAAG-3′
Sequence 2: 5′-AGUUAUUGAUGU-3′
Sequence 3: 5′-AGAACCCUUGUA-3′

(b)

Sequence 1: Met-Ala-Lys-Lys
Sequence 2: Ser-Tyr-(termination)
Sequence 3: Arg-Thr-Leu-Val

(c) The coding strand has the same sequence as the mRNA, except T is substituted for U:

5′-ATGGCAAAAAAG-3′

Solutions to Problems and Discussion Questions

2. Your essay should include a description of the nature and structure of the genetic code, the enzymes and logistics of transcription, and the chemical nature of polymerization.

4. The UUACUUACUUAC tetranucleotide sequence will produce the following triplets depending on the initiation point: UUA = Leu; UAC = Tyr; ACU = Thr; CUU = Leu. Notice that because of the degenerate code, two codons correspond to the amino acid leucine.

The UAUCUAUCUAUC tetranucleotide sequence will produce the following triplets depending on the initiation point: UAU = Tyr; AUC = Ile; UCU = Ser; CUA = Leu. Notice that in this case, degeneracy is not revealed, and all the codons produce unique amino acids.

6. Given that AGG = Arg, then information from the AG copolymer indicates that AGA also codes for Arg and GAG must therefore code for Glu. Coupling this information with that of the AAG copolymer, GAA must also code for Glu, and AAG must code for Lys.

8.

Original		*Substitutions*
threonine	----->	*alanine*
AC(U, C, A, or G)		GC(U, C, A, or G)
glycine	----->	*serine*
GG(U or C)		AG(U or C)
isoleucine	----->	*valine*
AU(U, C, or A)		GU(U, C, or A)

10. Polynucleotide phosphorylase generally functions in the degradation of RNA; however, in an *in vitro* environment, with high concentrations of the ribonucleoside diphosphates, the direction of the reaction can be forced toward polymerization. *In vivo*, the concentration of ribonucleoside diphosphates is low and the degradative process is favored.

12. Applying the coding dictionary, the following sequences are "decoded":

Sequence 1: Met-Pro-Asp-Tyr-Ser-(term)
Sequence 2: Met-Pro-Asp-(term)

The 12th base (a uracil) is deleted from Sequence 1, thereby causing a frameshift mutation, which introduced a terminating triplet UAA.

14.

G G A Gly (wild type)	G G U Gly (wild type)
U G A **term**	U G U **Cys**
C G A **Arg**	C G U **Arg**
A G A **Arg**	A G U **Ser**
G U A **Val**	G U U **Val**
G C A **Ala**	G C U **Ala**
G A A **Glu**	G A U **Asp**
G G U Gly	G G A Gly
G G C Gly	G G C Gly
G G A Gly	G G G Gly

16. The number of codons for each particular amino acid (synonyms) is directly related to the frequency of amino acid incorporation stated in the problem.

18. First, DNA does not directly participate in protein synthesis. DNA is located in the nucleus of a eukaryotic cell, whereas protein synthesis occurs in the cytoplasm. Second, RNA, which is chemically similar to DNA, is synthesized in the nucleus of eukaryotic cells. Much of the RNA migrates to the cytoplasm, the site of protein synthesis. Third, there is generally a direct correlation between the amounts of RNA and protein in a cell. More direct support was derived from experiments showing that an RNA other than that found in ribosomes was involved in protein synthesis.

20. Ribonucleoside triphosphates and a DNA template in the presence of RNA polymerase and a divalent cation (Mg^{2+}) produce a ribonucleoside monophosphate polymer, and release DNA, and pyrophosphate (diphosphate). Equimolar amounts of precursor ribonucleoside triphosphates and products (polymer and pyrophosphates) are formed. The polymer grows by sequential addition of ribonucleoside monophosphates, derived from ribonucleoside triphosphates, with the release of pyrophosphate. In *E. coli*, transcription and translation can occur simultaneously. Ribosomes add to the 5′ end nascent mRNA and progress to the 3′ end during translation.

22. An RNA transcript that is destined to become an mRNA often involves modification of the 5′ end, to which a 7-methylguanosine cap is added. In addition, a stretch of as many as 250 adenylic acid residues is often added to the 3′ end after removal of an AAUAAA sequence. The vast majority of eukaryotic pre-mRNAs also contain intervening sequences that are removed, often in a variety of combinations, during the maturation process. In some organisms, RNA editing occurs in one of two ways: substitution editing where nucleotides are altered and insertion/deletion editing that changes the total number of bases.

Processing location	*Example*
5′ end	Addition of 7-mG
3′ end	Poly-A addition
internal	Removal of internal sequences
	RNA editing
	Substitution
	Insertion/deletion

24. Substitution editing occurs when individual nucleotide bases are altered. It is very common in mitochondrial and chloroplast RNAs as well as some nuclear-derived eukaryotic RNAs. The protein apolipoprotein B occurs in a long and short form, even though a single gene encodes both forms. The initial transcript is edited, which generates a stop codon that terminates the polypeptide at about half its length. The other category is insertion/deletion editing. The parasite that causes African sleeping sickness uses insertion/deletion editing of its mitochondrial RNAs in forming the initiation codon that then places the remaining sequence in a proper reading frame.

26. (a) The first two bases in the triplet code are common to more codons than the last base. If fewer amino acids were used in earlier times, perhaps the first two bases were primarily involved.

(b) All the amino acids mentioned as primitive use guanine as the first base in each codon. We might therefore suppose that the GNN configuration was the starting point of the present-day code. Within the GNN format, the second base is used to distinguish among the amino acids mentioned as primitive.

(c) It is interesting to note that what are considered as the most primitive amino acids are GNN coded, while the later-arriving amino acids are U(U, A, G)(U, C) coded. It would seem that some phenomenon could explain the differences in codon structure between primitive and late-arriving amino acids. It is likely that the addition of new amino acids to the codon field was a slow and mutation-prone process. Did the addition of late-arriving amino acids displace earlier codon assignments, or were some of the late-arriving amino acids able to make use of codons that originally did not code for an amino acid? If this is the case, fewer and more restricted codon assignments would be available for late-arriving amino acids. Notice that each of the late-arriving amino acids has the same starting nucleotide as the present-day stop codons. Is it possible that what were once stop codons were used for late-arriving amino acids because this would be less disruptive to protein synthesis than displacing assignments at the time of their introduction? Much is left to be discovered regarding the structure of the genetic code.

28. Proline: C_3 and one of the C_2A triplets
Histidine: one of the C_2A triplets
Threonine: one C_2A triplet and one A_2C triplet
Glutamine: one of the A_2C triplets
Asparagine: one of the A_2C triplets
Lysine: A_3

30. (a,b) Use the code table to determine the number of triplets that code each amino acid; then construct a graph.

(c) There appears to be a weak correlation between the relative frequency of amino acid usage and the number of triplets for each.

(d) To continue to investigate this issue, one might examine additional amino acids in a similar manner. In addition, different phylogenetic groups use code synonyms differently. It may be possible to find situations in which the relationships are more extreme. One might also examine more proteins to determine whether such a weak correlation is stronger with different proteins.

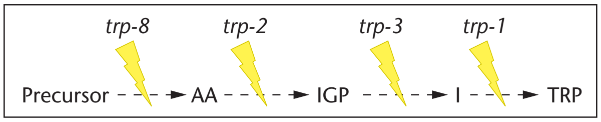
32. (a,b) Alternative splicing occurs when pre-mRNAs are spliced in more than one way to yield various combinations of exons in the final mRNA product. Upon translation of a group of alternatively spliced mRNAs, a series of related proteins, called isoforms, are produced. It is likely that alternative splicing provided an evolutionary advantage since a variety of functionally related proteins from one original source gene can be made in a particular tissue. In other words, varieties of similar proteins can be produced by alternative splicing rather than by independent evolution.

CHAPTER 14

Answers to Now Solve This

14.1 One can conclude that the amino acid is not involved in recognition of the codon.

14.2



14.3 There are two codons for glutamic acid: GAA and GAG. With two of the codons for valine being GUA and GUG, a single base change from glutamic acid's GAA or GAG could cause the Glu → Val switch. Likewise, single base changes to lysine's AAA or AAG could also cause a Glu → Lys switch. The normal glutamic acid is a negatively charged amino acid, whereas valine carries no net charge and lysine is positively charged. Given these significant charge changes, one would predict some, if not considerable, influence on protein structure and function. Such changes could stem from internal changes in folding (tertiary and/or quaternary structure) or from interactions with other molecules in the RBC, especially other hemoglobin molecules.

Solutions to Problems and Discussion Questions

2. When involved in protein synthesis, a ribosome will contain the following components: mRNA, charged tRNA, large and small ribosomal subunits, elongation and perhaps initiation factors, peptidyl transferase, GTP, Mg^{2+}, nascent proteins, and possibly GTP-dependent release factors. Together, these components order and form peptide bonds between adjacent amino acids, thereby forming proteins.

4. It was reasoned that there would not be sufficient affinity between amino acids and nucleic acids to account for protein synthesis. For example, acidic amino acids would not be attracted to nucleic acids. With an adaptor molecule, specific hydrogen bonding could occur between nucleic acids, and specific covalent bonding could occur between an amino acid and a nucleic acid tRNA.

6. Since there are three nucleotides that code for each amino acid, there would be 423 code letters (nucleotides), 426≈including a termination codon. This assumes that other features, such as the poly-A tail, the 5′ cap, and noncoding leader sequences, are omitted.

8. An amino acid in the presence of ATP, Mg^{2+}, and a specific aminoacyl synthetase becomes activated as an amino acid—AMP enzyme complex (+ PP_i). This complex interacts with a specific tRNA and transfers the activated amino acid to this tRNA to produce the aminoacyl tRNA.

10. Isoaccepting tRNAs are those tRNAs that recognize and accept only one type of amino acid. In some way, each of the 20 different aminoacyl tRNA synthetases must be able to recognize either the base composition and/or tertiary structure of each of the isoaccepting tRNA species. Otherwise, the fidelity of translation would be severely compromised. The most direct solution to the problem would be to have each synthetase recognize each anticodon. Another reasonable consideration might involve the variable loop, which, in conjunction with the anticodon, might enable such specificity. In reality, several characteristics of each tRNA are involved: one or more of the anticodon bases, portions of the acceptor arm, and a particular base that lies near the CCA terminus.

12. Phenylalanine is an amino acid that is required for protein synthesis. Whereas too much phenylalanine and its derivatives cause PKU in phenylketonurics, too little will restrict protein synthesis.

14. When an expectant mother returns to consumption of phenylalanine in her diet, she subjects her baby to higher than normal levels of phenylalanine throughout its development. Since increased phenylalanine is toxic, many (approximately 90 percent) newborns are severely and irreversibly retarded at birth. Expectant mothers (who are genetically phenylketonurics) should return to a low-phenylalanine intake during pregnancy.

16. The fact that enzymes are a subclass of the general term *protein*, a *one-gene:one-protein* statement might seem to be more appropriate. However, some proteins are made up of subunits, each different type of subunit (polypeptide chain) being under the control of a different gene. Under this circumstance, the *one-gene:one-polypeptide* statement might be more reasonable.

18. The electrophoretic mobility of a protein is based on a variety of factors, primarily the net charge of the protein and, to some extent, the conformation in the electrophoretic environment. Both are based on the primary structure of a protein. The interactions (hydrogen bonds) of the components of the peptide bonds, hydrophobic, hydrophilic, and covalent interactions (as well as others) are all dependent on the original sequence of amino acids and take part in determining the final conformation of a protein. A change in the electrophoretic mobility of a protein would therefore indicate that the amino acid sequence had been changed.

20. In the late 1940s, Pauling demonstrated a difference in the electrophoretic mobility of HbA and HbS and concluded that the difference had a chemical basis. Using the fingerprinting technique, Ingram determined that the chemical change occurs in the primary structure of the globin portion of the molecule. He found a change in the sixth amino acid in the β chain.

22. One would expect individuals with HbC to suffer some altered hemoglobin function and, perhaps, be resistant to malaria as well. In fact, HbC homozygotes suffer mild hemolytic anemia (a benign hemoglobinopathy). The HbC gene is distributed particularly in malarial-infested areas, suggesting that some resistance to malaria is conferred. Recent studies indicate that HbC may be protective against severe forms of malaria, but not to more uncomplicated forms.

24. *Primary*: the linear arrangement or sequence of amino acids. This sequence determines the higher-level structures.

Secondary: α-helix and β-pleated sheet structures generated by hydrogen bonds between components of the peptide bond.

Tertiary: folding that occurs as a result of interactions between the amino acid side chains. These interactions include, but are not limited to, the following: covalent disulfide bonds between cysteine residues, interactions of hydrophilic side chains with water, and interactions of hydrophobic side chains with each other.

Quaternary: the association of two (dimer) or more polypeptide chains. Called *oligomeric*, such a protein is made up of more than one protein chain.

26. Protein folding results in a thermodynamically favorable conformation. In folded proteins, polar groups tend to be found at the surface, where thermodynamically favorable interactions with surrounding water molecules can occur. Likewise, hydrophobic groups tend to be buried in the interior of folded proteins to avoid unfavorable interactions with water. Finally, disulfide bonds contribute to protein stability by covalently linking the side chains of two (possibly distant) cysteine residues that have been brought into close proximity by folding.

28. While there are many types of posttranslational modifications, your text lists the following six: Removal and/or modification of the N-terminal amino acid, modification of individual amino acids, attachment of carbohydrate side chains, trimming of polypeptide chains, removal of targeting sequence, coupling with prosthetic groups. In most cases, the modification serves to change the functionality of a protein (e.g., activation or inactivation of an enzyme).

30. Enzymes function to regulate catabolic and anabolic activities of cells. They lower the *energy of activation*, thus allowing chemical reactions to occur under conditions that are compatible with living systems. Enzymes possess active sites and/or other domains that are sensitive to the environment. The active site is considered to be a crevice, or pit, which binds reactants, thus enhancing their interaction. The other domains mentioned above may influence the conformation and, therefore, the function of the active site.

32. Even though three gene pairs are involved, notice that because of the pattern of mutations, each cross may be treated as (a) monohybrid or (b,c) dihybrid.

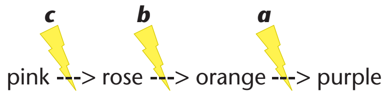
(a)	F_1: *AABbCC*	=	speckled	
	F_2: 3 *AAB_CC*	=	speckled	
	1 *AAbbCC*	=	yellow	
(b)	F_1: *AABbCc*	=	speckled	
	F_2: 9 *AAB_C_*	=	speckled	
	3 *AAB_cc*	=	green	
	3 *AAbbC*	=	yellow	} 4
	1 *AAbbcc*	=	yellow	
(c)	F_1: *AaBBCc*	=	speckled	
	F_2: 9 *A_BBC_*	=	speckled	
	3 *A_BBcc*	=	green	
	3 *aaBBC_*	=	colorless	} 4
	1 *aaBBcc*	=	colorless	

34. A cross of the following nature would satisfy the data:

AABBCC × *aabbcc*

Offspring in the F_2:

27	*A_B_C_*	=	purple
9	*A_B_cc*	=	pink
9	*A_bbC_*	=	rose
9	*aaB_C_*	=	orange
3	*A_bbcc*	=	pink
3	*aaB_cc*	=	pink
3	*aabbC_*	=	rose
1	*aabbcc*	=	pink



The above hypothesis could be tested by conducting a backcross as given below:

AaBbCc × *aabbcc*

The cross should give a 4 (pink) : 2 (rose) : 1 (orange) : 1 (purple) ratio.

36. (a) Since protein synthesis is dependent on the passage of mRNA from DNA to ribosomes, any circumstance that compromises this flow will cause a reduction in protein synthesis. An antisense oligonucleotide will bind to the target mRNA—the more specific the binding, the more specific the influence on protein synthesis. Ideally, a particular species of antisense oligonucleotide would affect one and only one mRNA and therefore only one protein population.

(b) A length of around 15 to 16 nucleotides is most effective in causing RNA degradation.

(c) The oligonucleotide must be able to enter the interior of target cells. Once inside, it needs to remain intact. It is likely that stability of the oligonucleotide is dependent on its base composition and length. The oligonucleotide must be small enough to diffuse effectively throughout the cell in order to “locate” the targeted mRNA, and it must not assume a folded conformation. It is also likely that the actual location of binding to the target is important in mRNA degradation.

CHAPTER 15

Answers to Now Solve This

15.1 The phenotypic influence of any base change is dependent on a number of factors, including its location in coding or noncoding regions, its potential in dominance or recessiveness, and its interaction with other base sequences in the genome. If a base change is located in a noncoding region, there may be no influence on the phenotype. However, some noncoding regions serve regulatory functions—mutations that influence transcription levels, polyA addition, splicing, and translation could affect phenotype. Furthermore, some noncoding regions in a traditional sense may influence other genes and/or gene products. If a mutation that occurs in a coding region acts as a full recessive, there should be no influence on the phenotype. If a mutant gene acts as a dominant, then there would be an influence on the phenotype. Some genes interact with other genes in a variety of ways that would be difficult to predict without additional information.

15.2 If a gene is incompletely penetrant, it may be present in a population and only express itself under certain conditions. It is unlikely that the gene for hemophilia behaved in this manner. If a gene's expression is suppressed by another mutation in an individual, it is possible that offspring may inherit a given gene and not inherit its suppressor. Such offspring would have hemophilia. It is possible that the mutation in Queen Victoria's family was a new germ line, arising in the father. Lastly, it is possible that the mother was heterozygous, and by chance, no other individuals in her family were unlucky enough to receive the mutant gene.

15.3 Any agent that inhibits DNA replication, either directly or indirectly, through mutation and/or DNA cross linking, will suppress the cell cycle and may be useful in cancer therapy. Since guanine alkylation often leads to mismatched bases, they can often be repaired by a variety of mismatched repair mechanisms. However, DNA cross linking can be repaired by recombinational mechanisms. Thus, for such agents to be successful in cancer therapy, suppressors of DNA repair systems are often used in conjunction with certain cancer drugs.

15.4 Ethylmethane sulfonate (EMS) alkylates the keto groups at the sixth position of guanine and at the fourth position of thymine. In each case, base-pairing affinities are altered and transition mutations result. Altered bases are not readily repaired, and once the transition to normal bases occurs through replication, such mutations avoid repair altogether.

Solutions to Problems and Discussion Questions

2. Your essay should include a brief description of the genomic differences between eukaryotes and bacteria and the ways that ploidy influences the phenotypic effects of mutations in one copy of a gene. You should also include a summary of repair pathways that operate predominantly in bacteria or predominantly in eukaryotes, as well as a description of the differences in repair pathways that are shared by both types of organisms.

4. A recessive somatic mutation would not produce a visible phenotype in a diploid organism. A dominant mutation would be more likely to be visible if it occurs early in development. If only a relatively small number of cells are affected, the effects of a dominant mutation could be masked by surrounding nonmutant cells. Similarly, if the mutation does not cause substantial alterations to the gene product or to regulation of expression of this product, an effect may not be visible. Those that occur in somatic cells are not transmitted to the next generation but may lead to altered cellular function or tumors.

6. A gene is likely to be the product of perhaps a billion or so years of evolution. Each gene and its product function in an environment that has also evolved, or coevolved. A coordinated output of each gene product is required for life. Deviations from the norm, caused by mutation, are likely to be disruptive because of the complex and interactive environment in which each gene product must function. However, on occasion a beneficial variation occurs.

8. A silent mutation is a point mutation in an open reading frame that does not alter the amino acid encoded, due to degeneracy of the genetic code. A neutral mutation is a mutation that occurs in noncoding DNA and does not affect gene products or gene expression.

10. All three of the agents are mutagenic because they cause base substitutions. Deaminating agents oxidatively convert an amino group to a keto group such that cytosine is converted to uracil and adenine is converted to hypoxanthine. Uracil pairs with adenine, and hypoxanthine pairs with cytosine. Alkylating agents donate an alkyl group to the amino or keto groups of nucleotides, thus altering base-pairing affinities. 6-ethyl guanine acts like adenine, thereby pairing with thymine. Base analogs such as 5-bromouracil and 2-amino purine are incorporated as thymine and adenine, respectively, yet they base-pair with guanine and cytosine, respectively.

12. X rays are of higher energy and shorter wavelength than UV light. They have greater penetrating ability and can create more disruption of DNA.

14. *Photoreactivation* can lead to repair of UV-induced damage. The photoreactivation enzyme, will absorb a photon of light to cleave thymine dimers. *Excision repair* involves the products of several genes, DNA polymerase I and DNA ligase, to clip out the UV-induced dimer, fill in, and join the phosphodiester backbone in the resulting gap. The excision repair process can be activated by damage that distorts the DNA helix. *Recombinational repair* is a postreplication repair system that responds to DNA that has escaped repair at the time of replication. If a gap is created when DNA polymerase stalls and skips replication on one of the newly synthesized strands, recombinational repair fills this gap by allowing genetic exchange with the undamaged template strand of the same polarity. The resulting gap on the donor strand is filled by repair synthesis. Finally, when a large amount of mismatches and gaps are detected, a "rescue operation or SOS response" is activated. Many different gene products are involved in this repair process: *recA* and *lexA*. In SOS repair, random or incorrect nucleotides are often incorporated at sites where DNA polymerase would normally stall; so this therefore is called an "error-prone system."

16. There are numerous regions upstream from coding regions in a gene that are sensitive to mutation. Many mutations upset the regions that signal transcription factor and/or polymerase binding, thereby influencing transcription. Mutations within introns may affect intron splicing or other factors that determine mRNA stability or translation.

18. *Xeroderma pigmentosum* is a rare recessive disorder in which affected individuals are highly sensitive to UV radiation and have a 2000-fold higher incidence of cancer than unaffected individuals. Cells from XP patients are unable to undergo unscheduled DNA synthesis, which is a step in the nucleotide excision repair system. Studies with heterokaryons provided evidence for at least seven different genes involved in the NER pathway. Since cancer is caused by mutations in several types of genes, interfering with DNA repair can enhance the occurrence of these types of mutations.

20. It is possible that through the reduction of certain environmental agents that cause mutations, mutation rates might also be reduced. On the other hand, certain industrial and medical activities actually concentrate mutagens (radioactive agents and hazardous chemicals). Unless human populations are protected from such agents, mutation rates might actually increase. If one asks about the accumulation of mutations (not rates) in human populations as a result of improved living conditions and medical care, then it is likely that as the environment becomes less harsh (through improvements), more mutations will be tolerated as selection pressure decreases. In addition, as individuals live longer and have children at a later age, some studies indicate that older males accumulate more gametic mutations.

22. Transposons cause changes in DNA in a variety of ways, including massive chromosomal alterations. In most cases, changes in DNA are harmful to organisms, while in rare cases, an evolutionary advantage occurs because the new genetic variation confers a selective advantage.

24.

XP1	*XP4*	*XP5*
XP2		*XP6*
XP3		*XP7*

These groupings (complementation groups) indicate that at least three "genes" form products necessary for unscheduled DNA synthesis. All of the cell lines that are in the same complementation group are defective in the same product.

26. Approximately 82 percent of humans' radiation exposure comes from natural sources. While diagnostic X-rays do contribute about 10 percent of the exposure, other human-made forms of radiation contribute only a relatively small amount. That's not to say that human-made radiation exposure is not a factor in causing mutations; rather, it is not a major factor.

28. Since Betazoids have a 4-letter genetic code and the gene is 3332 nucleotides long, the protein involved must be 832 amino acids in length (the last 4 nucleotides encode a termination codon).

(*mr-1*) Codon 829 specifies an amino acid that is very close to the 3′ end of the gene (the carboxyl terminus of the protein). Because a nonsense mutation would terminate translation prematurely, the protein would only be shortened by four amino acids. Thus, the protein's ability to fold and perform its cellular function must not be seriously altered. Because of the direction of translation (5′ to 3′ on the mRNA) the carboxyl terminal amino acids in a protein are the last to be included in folding priorities and are sometimes less significant in determining protein function.

(*mr-2*) Since the phenotype is mild, this amino acid change does not completely inactivate the protein, but it does change its activity to some extent. It is very possible that the substituted amino acid is chemically similar to the original and causes the protein to fold in a slightly aberrant manner, allowing it to have some residual function but preventing it from functioning entirely normally. In addition, even if the protein folds similarly to the wild-type protein, charge or structural differences in the protein's active site may be only mildly influenced.

(*mr-3*) This deletion contains a total of 68 nucleotides, which account for 17 amino acids. Since Betazoids' codons contain four nucleotides, the mRNA reading frames are maintained subsequent to the deletion. Protein function significantly depends on the relative positions of secondary levels of structure: α-helices and β-sheets. If the deleted section is a "benign" linker between more significant protein domains, then perhaps the protein can tolerate the loss of some amino acids in a part of the protein without completely losing its function.

(*mr-4*) The amino acid specified by codon 192 must be critical to the function of the protein. Altering this amino acid must disrupt a critical region of the protein, thus causing it to lose most or all of its activity. If the protein is an enzyme, this amino acid could be located in its active site and be critical for the ability of the enzyme to bind and/or influence its substrate. One might expect that the amino acid alteration is rather radical such as one sees in the generation of sickle-cell anemia. HbS is caused by the substitution of a valine (no net charge) for glutamic acid (negatively charged).

(*mr-5*) A deletion of 11 base pairs, a number that is not divisible by four, will shift the reading frames subsequent to its location. Even though this deletion is smaller than the deletion discussed above (83—150) and is located in the same region, it causes a reading frame shift. In addition, some or all of the amino acids that are added downstream from the mutation may be different from those in the normal protein. All will not likely change because of synonyms in the code. There is also the possibility that a nonsense triplet quadruplet may be introduced in the "out-of-phase" region, thus causing premature chain termination. Because this mutation occurs early in the gene, most of the protein will be affected. This may well explain the severe insensitive phenotype.

30. First, one might suggest that transposons, for one reason or another, are more likely to insert in noncoding regions of the genome. One might also suggest that they are more stable in such regions. Second, and more likely, it is possible that transposons insert randomly and that selection eliminates those that have interrupted coding regions of the genome. Since such regions are more likely to influence the phenotype, selection is more likely to influence such regions.

CHAPTER 16

Answers to Now Solve This

16.1 (a) Because of the deletion of a base early in the *lac Z* gene, there will be "frameshift" of all the reading frames downstream from the deletion, thereby altering many amino acids. It is likely that either premature chain termination of translation will occur (from the introduction of a nonsense triplet in a reading frame) or the normal chain termination will be ignored. Regardless, a mutant condition for the *Z* gene will be likely. If such a cell is placed on a lactose medium, it will be incapable of growth because β-galactosidase is not available.

(b) If the deletion occurs early in the *A* gene, one might expect impaired function of the *A* gene product, but it will not influence the use of lactose as a carbon source.

16.2 (a) With no lactose and no glucose, the operon is off because the *lac* repressor is bound to the operator, and although CAP is bound to its binding site, it will not override the action of the repressor.

(b) With lactose added to the medium, the *lac* repressor is inactivated and the operon is transcribing the structural genes. With no glucose, the CAP is bound to its binding site, thus enhancing transcription.

(c) With no lactose present in the medium, the *lac* repressor is bound to the operator region, and since glucose inhibits adenyl cyclase, the CAP protein will not interact with its binding site. The operon is therefore "off."

(d) With lactose present, the *lac* repressor is inactivated; however, since glucose is also present, CAP will not interact with its binding site. Under this condition transcription is severely diminished, and the operon can be considered to be "off."

Solutions to Problems and Discussion Questions

2. Your essay should include a description of the evolutionary advantages of the efficient response to environmental resources and challenges (antibiotics, for example) when such resources are present and likewise, the ability to turn off metabolic functions when they are not needed. Having related functions in operons provides for coordinated responses.

4. In an *inducible system*, the repressor that normally interacts with the operator to inhibit transcription is inactivated by an *inducer*, thus permitting transcription. In a *repressible system*, a normally inactive repressor is *activated* by binding a *co-repressor*. The activated repressor will then bind to the operator to inhibit transcription. Because the interaction of the protein (repressor) has a negative influence on transcription, the systems described here are forms of *negative control*.

6. $I^+O^+Z^+$ = Because of the function of the active repressor from the I^+ gene, and no lactose to influence its function, there will be **no enzyme made**.

$I^+O^CZ^+$ = There will be a **functional enzyme made** because the constitutive operator is in *cis* with a *Z* gene. The lactose in the medium will have no influence because of the constitutive operator. The repressor cannot bind to the mutant operator.

$I^-O^+Z^-$ = There will be a **nonfunctional enzyme made** because with I^- the system is constitutive, but the *Z* gene is mutant. The absence of lactose in the medium will have no influence because of the nonfunctional repressor. The mutant repressor cannot bind to the operator.

$I^-O^+Z^-$ = There will be a **nonfunctional enzyme made** because with I^- the system is constitutive, but the *Z* gene is mutant. The lactose in the medium will have no influence because of the nonfunctional repressor. The mutant repressor cannot bind to the operator.

$I^-O^+Z^+/F'I^+$ = There will be **no enzyme made** because in the absence of lactose, the repressor product of the I^+ gene will bind to the operator and inhibit transcription.

$I^+O^CZ^+/F'O^+$ = Because there is a constitutive operator in *cis* with a normal *Z* gene, there will be a **functional enzyme made**. The lactose in the medium will have no influence because of the mutant operator.

$I^+O^+Z^-/F'I^+O^+Z^+$ = Because there is lactose in the medium, the repressor protein will not bind to the operator and transcription will occur. The presence of a normal *Z* gene allows a **functional and nonfunctional enzyme to be made**. The repressor protein is diffusible, working in *trans*.

$I^-O^+Z^-F'I^+O^+Z^+$ = Because there is no lactose in the medium, the repressor protein (from I^+) will repress the operators and there will be **no enzyme made**.

$I^sOI^SO^+Z^+/F'O^+$ = With the product of I^S there is binding of the repressor to the operator and therefore **no enzyme made**. The lack of lactose in the medium is of no consequence because the mutant repressor is insensitive to lactose.

$I^+O^CZ^+/F'O^+Z^+$ = The arrangement of the constitutive operator (O^C) with the *Z* gene will cause a **functional enzyme to be made**.

8. A single *E. coli* cell contains very few molecules of the *lac* repressor. However, the *lac* I^q mutation causes a 10 × increase in repressor protein production, thus facilitating its isolation. With the use of dialysis against a radioactive gratuitous inducer (IPTG), Gilbert and Müller-Hill were able to identify the repressor protein in certain extracts of *lac* I^q cells. The material that bound the labeled IPTG was purified and shown to be heat labile and have other characteristics of protein. Extracts of *lac* I^- cells did not bind the labeled IPTG.

10. (a) Because activated CAP is a component of the cooperative binding of RNA polymerase to the *lac* promoter, absence of a functional *crp* would compromise the positive control exhibited by CAP.

(b) Without a CAP binding site there would be a reduction in the inducibility of the *lac* operon.

12. Attenuation functions to reduce the synthesis of tryptophan when it is in full supply. It does so by reducing transcription of the *tryptophan* operon. The same phenomenon is observed when tryptophan activates the repressor to shut off transcription of the *tryptophan* operon.

14. Apparently, relatively high levels of neelaredoxin are produced at all times (*constitutively expressed*) even when potential inducers of gene expression are not added to the system. Additional neelaredoxin gene expression is not responsive (*induced*) as a result of O_2 and H_2O_2 treatment.

16. Attenuation of the *trp* operon in *E. coli* involves the amino acid tryptophan and the leader region of mRNA to form a termination configuration for transcription. Riboswitches encompass a variety of mechanisms that both terminate and allow transcription, depending on the particular metabolic requirements of the organism. In both cases, the formation of *intra*molecular double-stranded RNA induces conformational changes in the leader regions of mRNA resulting in gene regulation at the transcriptional level. sRNAs, by contrast, are separate transcripts, complementary to the message, that function through the formation of *inter*molecular double-stranded RNA. Regulation is achieved at the level of translation with binding of the sRNA to the message either preventing or enhancing translation.

18. Since a substance supplied in the medium (the antibiotic) causes the synthesis of the efflux pump components, two situations seem appropriate. Under a *negative control* system, the antibiotic would interrupt the repressor to bring about induction (this would be an inducible system). Under *positive control,* the antibiotic would activate an activator (this again would be an inducible system).

20. The regulatory gene product is exerting *positive control*.

(a) In wild-type cells, when tis is present, no enzymes are made; therefore, tis must inactivate the positive regulatory protein. When tis is absent, the regulatory protein is free to exert its positive influence on transcription.

(b) Mutations in the operator negate the positive action of the regulator.

22. (a) Call one type of constitutive mutation *lexA*$^-$ (mutation in the repressor gene product) and the other O^{uvrA-} (mutation in the operator).

(b) One can make partial diploid strains using F′. O^{uvrA-} will (given the other genes brought in by the F′ element) be dominant to O^{uvrA+} and *lexA*$^-$ will be recessive to *lexA*$^+$. O^{uvrA-} will act in *cis*.

24. You will need to identify the complementary regions. You will find four regions that "fit."

CHAPTER 17

Answers to Now Solve This

17.1 Cancer cells often originate under the influence of mutations in tumor-suppressor genes or proto-oncogenes. Should hypermethylation occur in one of many DNA repair genes, the frequency of mutation would increase because the DNA repair system is compromised. The resulting increase in mutations might occur in tumor-suppressor genes or proto-oncogenes.

17.2 General transcription factors associate with a promoter to stimulate transcription of a specific gene. Some *trans*-acting elements, when bound to enhancers, interact with coactivators to enhance transcription by forming an enhanceosome that stimulates transcription initiation. Transcription can be repressed when certain proteins bind to silencer DNA elements and generate repressive chromatin structures. The same molecule may bind to a different chromosomal regulatory site (enhancer or silencer), depending on the molecular environment of a given tissue type.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of chromatin structure, the general architecture and functions of regulatory elements within chromatin, the activities performed by activators and chromatin modifiers, and the consequences of these activities.

4. Chromatin remodeling is one method of nucleosome modification, in which remodeling complexes reposition or reconfigure nucleosomes using ATP as an energy source. Examples are the replacement of histone H2A with the histone variant H2A.Z and the removal of nucleosomes bound to the *GAL* promoters by the chromatin remodeling complex, SWI/SNF.

6. When DNA is transcriptionally active, it is in a less condensed state and as such, more susceptible to DNase digestion.

8. Remodeling complexes are recruited to promoters by activators or repressors to either open or close the promoter. Chromatin remodeling complexes alter chromatin conformation by swapping histone variants, repositioning nucleosomes along the DNA, or pulling the DNA from the core nucleosome. Alterations that create a more open chromatin conformation will allow transcription, whereas downregulation of transcription will result if a more closed conformation is created.

10. Positively charged histones are able to interact with negatively charged phosphate groups of the DNA backbone. Addition of acetyl groups to histone tails weakens this interaction by reducing the positive charge on histones.

12. The effects of enhancers are limited by insulators—protein-binding sites found between the enhancer and promoter of a non-target gene. Insulators likely work by binding proteins that induce DNA looping that favors enhancer interactions with target promoter and blocks interactions with non-target promoters.

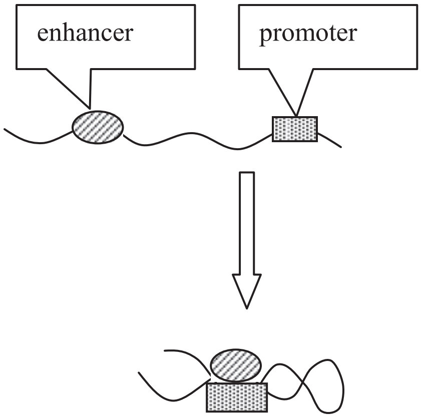
14. *Similarities:* Transcription initiation requires interaction between *cis*-acting elements and the *trans*-acting factors. Promoters are required elements and are located upstream of the transcribed gene. Activators and repressors can influence promoter recognition. The formation of DNA loops (although of different structure) contributes to regulation of transcription initiation.

Differences: In bacteria, the promoter is recognized by the σ subunit of the RNA polymerase holoenzyme. Different σ subunits recognize different promoter sequences, thereby regulating transcriptional specificity. Repressor proteins that induce DNA conformational changes (repression loops) can prevent promoter binding by RNA polymerase. In addition, RNAs (attenuators, riboswitches, sRNAs) can either allow or repress initiation, making transcription responsive to environmental or cellular conditions. In eukaryotes, chromatin structure may need to be modulated to make the promoter more or less accessible to the transcriptional machinery. RNAP II is recruited to promoters by general transcription factors. In addition, activators and repressors bind to enhancers and silencers, respectively. It has been proposed that large DNA loops are induced, bringing promoters and enhancers (or silencers) close to each other.

16. *Focused* promoters appear to define transcription initiation at a single nucleotide, whereas *dispersed* promoters direct initiation from a number of nucleotides. Most genes of lower eukaryotes use focused transcription, and in general are genes that are highly regulated. Dispersed promoters are more often associated with genes that act constitutively.

18. The DBD allows the activator to bind to enhancer elements, to correctly position the protein to interact with the correct gene sequence. The AD allows the activator to interact with other proteins, such as coactivators, important to the formation of the pre-initiation complex. It would make sense for the same types of domains to exist in repressor proteins as well.

20. Following is a diagram of a possible mechanism by which supercoiling may positively influence enhancer activity over a relatively long distance.



22. (a) This would remove the DNA-binding domain and not allow transcriptional activation.

(b) With no Gal3p, there would be no disruption of the Gal4p/Gal80p complex and no transcription of the *GAL1* gene.

(c) If the Gal80p can't interact with Gal3p, the block of the Gal4p AD by Gal80p cannot be alleviated, so there would be no *GAL1* transcription.

(d) This would reduce transcription of the *GAL1* gene.

(e) Generally, mutations in the TATA box of a promoter reduce transcription of the relevant gene.

24. γ globin is expressed during fetal development, but becomes silenced, which can be accomplished by methylation of cytidine in a CpG island in or near the promoter. 5-azacytidine is an analog of cytidine that cannot be methylated. When incorporated into DNA, it stimulates the expression of genes.

(b) 5-azacytidine is not gene specific, so it is likely to have widespread influence on the genome, which could constitute a considerable health hazard.

26. Methylation outside of the transcriptional unit results in a small (less than 2-fold) reduction in luciferase expression, whereas methylation within the unit causes a drastic reduction in expression (up to 98-fold). The effect of methylation of the transcriptional unit may be enhanced by methylation outside the unit.

28. The mutation is likely in a tissue-specific enhancer element. Fish with pelvic fins possess a wild-type copy of the enhancer and can upregulate *Pitx1* expression in the pelvic region. Fish without pelvic fins possess the mutated enhancer. While these fish are unable to upregulate expression in the pelvic area, expression in the jaw and pituitary gland remains normal.

30. The description of the regulatory element (location, effect on transcription upon mutation) suggests that it might be an enhancer. Interestingly, Gallagher et al. describe a different element—a barrier insulator—in this region. The insulators described in the text were those that prevented an enhancer from upregulating expression from a non-target gene. A barrier insulator prevents the spreading of heterochromatin into regions of euchromatin so that silencing of non-target genes does not occur. A mutation in a barrier insulator would allow heterochromatin to spread and would result in inappropriate gene silencing, as was seen for the *ANK1* gene.

CHAPTER 18

Answers to Now Solve This

18.1 A homozygous null mutation would produce XX males, and a constitutive *tra* mutation would produce XY females.

18.2 RISC searches for target mRNAs, guided by single-stranded miRNA molecules. If the miRNA and message are perfectly matched, RISC cleaves the target mRNA in the middle of the double-stranded sequence, leading to message degradation. In the case where the miRNA is a partial match, translation is blocked.

18.3 Without the zip code sequence, the actin mRNA would not be bound by ZPB1 and, consequently, neither silenced nor transported. As a result, actin would be produced in the cytoplasm near the nucleus and not in the lamellipodium.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of each type of regulation, the nature of each type of *cis* element, the action of the RBP, and the consequence of this action.

4. The use of alternative polyadenylation for the primary transcript of the CT/CGRP gene results in production of two mRNAs. In the message encoding the CT protein, exon 4, which contains a polyadenylation signal, is present, whereas in the message encoding the CGRP protein, exon 4 is skipped and the polyadenylation signal from a downstream exon is used.

6. *Cis*-acting sequences that promote (splicing enhancers) or inhibit (splicing silencers) splicing are recognized by different classes of RNA-binding proteins (RBPs). Since expression of RBPs is often tissue-specific, alternative splicing can also be tissue-specific.

8. Normally, stop codons are located near the poly-A tail of the message or downstream of exon–exon junctions. Termination signals that do not meet these criteria are considered by the cell as premature.

10. P bodies are complexes that can accumulate messages that are not being actively translated. Because they contain decapping enzymes and exoribonucleases, P bodies are believed to be message decay centers, although some data suggest that they serve as message storage centers.

12. In general terms, a cytoplasmic protein, Dicer, processes double-stranded small noncoding RNA (sncRNA) molecules to produce shorter dsRNAs. These associate with RISC, where an Argonaut-family protein cleaves and discards one of the two strands. The retained strand guides

RISC to the complementary target message, where the complex acts to prevent expression. The specific mechanism of silencing depends on the type of sncRNA used.

14. Since miRNA must base-pair with its potential target to direct RISC the correct message, you can determine the sequence that is complementary to the miRNA and search for that sequence among the cellular mRNAs, whose sequences are also known.

16. ceRNAs are long noncoding RNAs that contain miRNA response elements (MREs). MREs are sequences that are complementary to miRNA and that serve as miRNA binding sites. When present, ceRNAs will compete with mRNA targets for binding of miRNA molecules, rendering the miRNA less effective.

18. Translation is often inactivated by preventing the interactions between the 5′ and 3′ ends necessary for initiation. For messages that contain the CPE (cytoplasmic polyadenylation element) *cis*-regulatory sequence, the cytoplasmic polyadenylation element binding protein (CPEB) recruits poly-A ribonuclease (PARN) to the mRNA tail. PARN drastically reduces the length of the poly-A tail, resulting in the binding of an insufficient number of poly-A binding proteins (PABPs) to support a stable interaction with initiation factors. Second, CPEB recruits the protein Maskin, which prevents interactions between initiation factors at the 5′ end of the message. Messages inactivated in this way are stored in complexes known as ribonucleoprotein particles. Upon receiving a signal to resume translation, CPEB is phosphorylated, resulting in a conformational change that releases PARN. A cytoplasmic poly-A polymerase lengthens the message tail, which is bound by multiple PABPs. These are able to displace Maskin and allow the formation of an initiation complex.

20. Even though a particular species of mRNA may be fairly uniformly distributed throughout a cell, it does not follow that it is uniformly translated. It is likely that different domains reside in cells that contain different translational signals. If an mRNA finds itself in a particular molecular environment, it may be destined for translation, whereas that same mRNA in another part of a cell may not have the environmental stimulation necessary for translation.

22. Ubiquitin ligases function to mark proteins for degradation. They do this by processive addition of the protein ubiquitin to lysine residues in target proteins as well as to lysine residues in the ubiquitin itself. Proteosomes recognize polyubiquinated proteins as substrates and digest them into small peptides.

24. You should run parallel sets of assays—your control will test a GFP construct without any *cis*-elements and your experimental samples will each test a GFP construct containing a *cis*-element of interest. Monitor fluorescence in all cells tested. An absence, change in the subcellular localization, decreased or extended longevity, etc., of the fluorescent signal as compared with the control will allow the determination of the potential function of each *cis*-element tested.

26. The observation that the mutation affects only premenopausal women suggests that estrogen plays a role in this splicing. A splice enhancer is a likely site for the mutation, since it interacts with binding proteins. It is possible that the conformation/function of one or more of these proteins is influenced by the presence of estrogen. If the binding affinity between the enhancer and protein is slightly reduced by the SNP, splicing may not be altered. However, if high levels of estrogen further reduce protein affinity for the enhancer or reduce the amount of binding protein made, splicing would be compromised, but only in premenopausal women.

28. If the MREs were located in a coding region, the sequence would be translated. This would require either that a unique MRE exist for each targeted mRNA or that mRNAs targeted by the same miRNA encode a conserved protein domain. As it is, the location of MREs in noncoding sequences allows miRNAs to be more versatile in their targeting.

30. Actin mRNA is bound by ZBP1, which prevents translation and interacts with cytoskeleton motor proteins to achieve the transport of the mRNA to lamellipodia. Phosphorylation of ZBP1 by activated Src frees the actin mRNA and leads to the synthesis of actin monomers and their polymerization into actin fibers. This allows forward migration of the cell. To migrate away from a repulsive signal, the cell would need to form lamellipodia on the face opposite the signal and would need some means of transducing the signal from the receiving face to the responding face in order to activate Src and trigger the synthesis and polymerization of actin.

CHAPTER 19

Answers to Now Solve This

19.1 Gene activation or silencing requires that the chromatin containing the gene and its regulatory regions be either “open” or “closed,” respectively. If these chromatin structures are to persist through multiple cell divisions, histones in the immediate vicinity of the gene (in *cis*) would need to carry the appropriate modifications.

19.2 The observation that decreasing methylation levels in low-MN adult rats alleviated the effects of poor nurturing suggests that the key may be DNA methylation. Check the methylation status of rat gametes, newborn pups, and nurtured pups from both low-MN and high-MN parents. Compare patterns of gametes to pups within each category to determine whether there is a difference. Also compare gamete patterns and pup patterns between categories. Are the gametic patterns the same as or different from the newborn pattern or the nurtured pup pattern? Do newborn pups show the same pattern as nurtured pups? A slightly different approach would be to ask whether the pattern of a pup born to a low-MN mother can be altered if nurtured by a high-MN mother (and *vice versa*).

Solutions to Problems and Discussion Questions

2. Your essay should include a description of the types of epigenetic changes that have been associated with cancers, which categories of genes are involved, and the consequences of the epigenetic changes. Explain how these changes initiate and maintain cancerous growth.

4. In general, periodic methylation occurs at CpG-rich regions and promoter sequences. When a gene is imprinted by methylation, it remains transcriptionally silent. In a mammalian embryo, imprinting may silence only the paternal set of chromosomes, for example. The majority are found in heterochromatin, such as the centromere, where it promotes chromosome stability and helps prevent replication and transposition of LINEs and SINEs.

6. Several groups of proteins are involved in histone modification. Some proteins add chemical groups to histones, others interpret modifications, and some proteins remove the added chemical groups. Such modifications influence the structure of chromatin by altering the accessibility of nucleosomes. These chromatin alterations "open" or "close" genes for transcription.

8. DNA methylation and histone alterations work in concert. When DNA is unmethylated and histones are acetylated, nucleosomes are spaced in the open configuration and transcription can occur. When DNA is methylated and histones are deacetylated, nucleosomes are relatively close together and transcription is suppressed.

10. MicroRNAs play a significant role in the developing embryo. They are involved with RNA silencing through RISCs that act as repressors of gene expression by making mRNAs less likely to be translated.

12. The histone code describes the patterns of reversible modifications of histones and the interactions between and among them that contribute to the regulation of gene expression.

14. When mutations occur in the imprinting of genes, called epimutations, heritable changes in gene activity may occur. Imprinting defects cause Prader-Willi syndrome, Angelman syndrome, Beckwith-Wiedemann syndrome, and several others. In most cases imprinted genes encode growth factors or genes that regulate growth factors.

16. While data are scant, some studies have shown that children born after *in vitro* fertilization are at risk for low to very low birth weight that may have resulted from abnormal imprinting. There also appears to be an increased risk of a child conceived via ART having Beckwith-Wiedemann syndrome, Prader-Willi syndrome, and Angelman syndrome. Children conceived by IVF that have one of these syndromes have reduced levels or loss of maternal-specific imprinting. Given these data, it would seem reasonable that such information should be provided to prospective ART users. Each couple would need to reach a decision based on available science and their own value and belief sets.

18. Mutations in HAT genes would prevent the acetylation of histones and, therefore, promote a "closed" chromatin configuration, preventing gene expression. This can result in the silencing of tumor-suppressor genes and the unregulated proliferation of the cell.

20. When the promoter methylation of the glucocorticoid receptor (GR) is high, GR expression is low and the animals are less able to adapt to stress. Conversely, when the methylation level at the GR promoter is low, GR expression is high and the animals are stress adaptive.

22. Modifications of lysine 4, lysine 9, lysine 14, lysine 27, lysine 36, and lysine 79 cause activation of gene expression. Modifications of lysine 9, serine 10, lysine 27, and lysine 79 cause repression of gene expression.

(a) Yes, modification of lysine residues 9, 27, and 79 can cause either activation or repression of expression.

(b) This can be resolved, for the most part, by taking into account the type of modification performed. The exception to this is third methylation of lysine 79, which can result in either activation or repression.

(c) The consequences of the third methylation of lysine 79 may depend on the interaction of this modified amino acid with other modified residue.

24. The environment of the histone modification is important in determining the outcome of the modification. The different outcomes of methylation of H3K9 alone versus in the presence of methylated H3K4 and H4K20 suggests there is an interaction among the three residues that leads to gene activation.

CHAPTER 20

Answers to Now Solve This

20.1 (a) To select for cells that have incorporated a recombinant plasmid, you would add tetracycline to the medium because the gene that confers resistance to tetracycline is intact in the recombinant plasmid. Bacteria that have been transformed with the recombinant plasmid will, therefore, be resistant to tetracycline.

(b) Colonies that grow on a tetracycline medium but do not grow on the ampicillin medium probably contain the *Drosophila* DNA insert.

(c) Resistance to both antibiotics by a transformed bacterium could be explained in several ways. First, if cleavage with the *Pst*I was incomplete, then no change in biological properties of the uncut plasmids would be expected. Also, it is possible that the cut ends of the plasmid were ligated together in the original form with no insert.

20.2 Using the human nucleotide sequence, identify regions of the β-globin gene that are relatively conserved among mammals. Select sequences from these regions to create PCR primers, and amplify the sequences. One can then sequence the amplified DNA and compare the nucleotide and deduced amino sequences against the human counterpart. One can also use the amplified DNA to produce a probe to screen the library of the African okapi. The probe will hybridize to complementary sequences in the ocampi genomic library and identify the library clone containing the DNA of interest. Cells with the desired clone are then picked from the original plate, and the plasmid is isolated from the cells.

Solutions to Problems and Discussion Questions

2. Your essay should include an appreciation for the relative ease with which sections of DNA can be inserted into various vectors and the amplification and isolation of such DNA. You should also include the possibilities of modifying recombinant molecules.

4. Even though the human gene coding for insulin contains a number of introns, a cDNA generated from insulin mRNA is free of introns. Plasmids containing insulin genes (from cDNA) are free of introns, so no processing issue surfaces.

6. This segment contains the palindromic sequence of GGATCC, which is recognized by the restriction enzyme *Bam*HI. It also contains the sequence of GATC, which is recognized by the restriction enzyme *Sau*3AI. The double-stranded sequences are shown here:

GGATCC	GATC
CCTAGG	CTAG
*Bam*HI	*Sau*3AI

8. Plasmids were the first to be used as cloning vectors, and they are still routinely used to clone relatively small fragments of DNA. Because of their small size, they are relatively easy to separate from the host bacterial chromosome, and they have relatively few restriction sites. They can be engineered fairly easily (i.e., polylinkers and reporter genes added). Plasmids suffer from the limitation that they can only use bacteria as hosts. BACs are artificial bacterial chromosomes that can be engineered for certain qualities and are able to accept substantially larger DNA inserts.

YACs (yeast artificial chromosomes) contain telomeres, an origin of replication, and a centromere and are extensively used to clone DNA in yeast. With selectable markers (TRP1 and URA3) and a cluster of restriction sites, DNA inserts ranging from 100 kb to 1000 kb can be cloned and inserted into yeast. Since yeast, being a eukaryote, undergoes many of the typical RNA and protein processing steps of other, more complex eukaryotes, the advantages are numerous when working with eukaryotic genes.

10. Scientists were concerned about the consequences of the unintended release of recombinant DNA and/or any organisms modified by such DNA. Concerns were also raised about possible risks of combining eukaryotic and bacterial DNA. While greater understanding of the technology has allayed many of these concerns, some still persist. The most prevalent today is concern about foods containing products from genetically modified plants and animals.

12. The total number of molecules after 15 cycles would be 16,384, or $(2)^{14}$.

14. A cDNA library provides DNAs from RNA transcripts and is, therefore, useful in identifying what are likely to be functional DNAs. If one desires an examination of noncoding as well as coding regions, a genomic library would be more useful.

16. Several factors might contribute to the lack of representation of the 5′ end of the mRNA. One has to deal with the possibility that the reverse transcriptase may not completely synthesize the DNA from the RNA template. The other reason may be that the 3′ end of the copied DNA tends to fold back on itself, thus providing a primer for the DNA polymerase. Additional preparation of the cDNA requires some digestion at the folded region. Since this folded region corresponds to the 5′ end of the mRNA, some of the message is often lost.

18. Taking the number of bases recognized by *Bam*HI as 6, there would be approximately 4096 base pairs between sites. Given that λ DNA contains approximately 48,500 base pairs, there would be about 11.8 sites (48,500/4096).

20. *Taq* polymerase is isolated from a bacterium called *Thermus aquaticus*, which typically lives in hot springs. It is a heat-stable enzyme that can tolerate extreme temperature changes.

22. FISH involves the hybridization of a labeled probe to a complementary stretch of DNA in a chromosome. As such, it can be used to locate a specific DNA sequence (often a gene or gene fragment) in a chromosome. Spectral karyotyping uses FISH to detect individual chromosomes, a distinct advantage in identifying chromosomal abnormalities.

24. If a transgene integrates into a coding or regulatory region, it will likely cause a mutation and, as a result, change the phenotype of the organism in an unexpected fashion. Since two different genetic events occurred (disruption of one gene by addition of another), it could be difficult to tease apart the effects of the one event from those of the other.

26. Reporter genes, surface traits, various probes, and amplification by PCR are often used to determine the integration of a transgene.

28. One concern that has already been raised is the potential of creating so-called designer babies, whose genomes are edited for nonmedical reasons. Limiting use of CRISPR-Cas to edit only certain human genes would raise a number of ethical questions. The ethics of genome editing will certainly need to be discussed, and no doubt, guidelines will need to be established.

30. $T_m(°\mathrm{C}) = 81.5 + 0.41(\%\mathrm{GC}) - (675/N) = 81.5 + 0.41(33.3) - (675/21) =$ about 63°C. Subtracting 5°C gives us a good starting point of about 58°C for PCR with this primer. Notice that as the % of GC and length increase, the $T_m(°\mathrm{C})$ increases. GC pairs contain three hydrogen bonds rather than two as between AT pairs.

32. (a) As the percent dT increases, the required annealing temperature decreases because there are only two hydrogen bonds stabilizing AT base pairs.

(b) Each primer in a pool of random sequence primers will have a different base composition, so many will have different melting temperatures. This will complicate the determination of the proper annealing temperature to use in the experiment.

(c) One could isolate RNAs from each side, trap the mRNAs using oligo(dT), and analyze the mRNAs using quantitative PCR.

34. By examining previous equations, it is clear that a variety of factors influence T_m and, therefore, annealing of primers with DNA. Primers with different percentages of GC and/or length will have different annealing temperatures. To factor these variables into a single T_m given a number of other factors (divalent and monovalent cation concentrations, primer and target concentrations) has, to date, been a complex problem that is often unresolvable.

CHAPTER 21

Answers to Now Solve This

21.1 (a) To annotate a gene, one identifies gene-regulatory sequences found upstream of genes (promoters, enhancers, and silencers), downstream elements (termination sequences), and in-frame triplet nucleotides that are part of the coding region of the gene. In addition, 5′ and 3′ splice sites that are used to distinguish exons from introns as well as polyadenylation sites are also used in annotation.

(b) Similarity to other annotated sequences often provides insight as to a sequence's function and may serve to substantiate a particular genetic assignment. Direct sequencing of cDNAs from various tissues and developmental stages aids in verification.

(c) The 3141 genes identified on chromosome 1 constitute 15.7 percent of the total number of genes in the human genome (estimated to be 20,000). Since chromosome 1 contains 8 percent of the human genome and almost 16 percent of the genes, it would appear that chromosome 1 is gene rich.

21.2 Since structural and chemical factors determine the function of a protein, it is likely that several proteins share a considerable amino acid sequence identity. Since the *in vivo* function of such a protein is determined by secondary and tertiary structures, as well as local surface chemistries in active or functional sites, the nonidentical sequences may have considerable influence on function. Note that the query matches to different site positions within the target proteins. A number of other factors that suggest different functions include associations with other molecules (cytoplasmic, membrane, or extracellular), chemical nature and position of binding domains, posttranslational modification, signal sequences, and so on.

21.3 Because blood is relatively easy to obtain in a pure state, its components can be analyzed without fear of tissue-site contamination. Second, blood is intimately exposed to virtually all cells of the body and may therefore carry chemical markers of certain abnormal cells. It represents, theoretically, an ideal probe into the human body. However, when blood is removed from the body, its proteome changes, and those changes are dependent on a number of environmental factors. Thus, what might be a valid diagnostic for one condition might not be so for other conditions. In addition, the serum proteome is subject to change depending on the genetic, physiologic, and environmental state of the patient. Age and sex are additional variables that must be considered. Validation of a plasma proteome for a particular cancer would be strengthened by demonstrating that the stage of development of the cancer correlates with a commensurate change in the proteome in a relatively large, statistically significant pool of patients. Second, the types of changes in the proteome should be reproducible and, at least until complexities are clarified, involve tumorigenic proteins. It would be helpful to have comparisons with archived samples of each individual at a disease-free time.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of traditional recombinant DNA technology, which involved cutting and splicing genes, as well as descriptions of modern methods of synthesizing genes of interest, PCR amplification, microarray analysis, and so forth.

4. Whole-genome shotgun sequencing involves randomly cutting the genome into numerous smaller segments. Overlapping sequences are used to identify segments that were once contiguous, eventually producing the entire sequence. Difficulties in alignment often occur in repetitive regions of the genome. Map-based sequencing relies on creation of restriction maps of a chromosome, cloning into BACs or YACs, and further subcloning into smaller vectors prior to sequencing. After sequencing, similarly to the WGS method, alignment overlaps are identified and used to assemble a chromosome. Compared to whole-genome sequencing, the map-based approach is somewhat cumbersome and time consuming. Whole-genome sequencing has become the most common method for assembling genomes, with map-based cloning being used to resolve the problems often encountered during whole-genome sequencing.

6. One usually begins to annotate a sequence by comparing it to known sequences. Similarity to other annotated sequences often provides insight as to a sequence's function. Hallmarks to annotation are the identification of gene regulatory sequences found upstream of genes (such as promoters), downstream elements (termination sequences), and triplet nucleotides that are part of the coding region of the gene. Bacterial genes do not contain a number of the elements found in eukaryotic genes, so their annotation is sometimes less complicated. In eukaryotes, upstream elements would also include enhancers and silencers and downstream elements would also include a polyadenylation signal sequence. In addition, 5′ and 3′ splice sites that distinguish exons from introns are also used in annotation.

8. One initial approach to annotating a sequence is to compare the newly sequenced genomic DNA to known sequences already stored in various databases. The National Center for Biotechnology Information (NCBI) provides access to BLAST (Basic Local Alignment Search Tool) software, which directs searches through databanks of DNA and protein sequences. A segment of DNA can be compared to sequences in major databases such as GenBank to identify matches that align in whole or in part. For example, using a query sequence from mouse chromosome 11, one might find identical or similar sequences in a number of taxa. BLAST will compute a similarity score or identity value to indicate the degree to which two sequences are similar, as well as an expect value (E-value, the likelihood that the sequence matches by chance) that indicates the level of significance of a match (a value close to 1 indicates the match may be random). BLAST is one of many sequence alignment algorithms (RNA—RNA, protein—protein, etc.) that may sacrifice sensitivity for speed.

10. The main goals of the Human Genome Project are to establish, categorize, and analyze functions for human genes. As stated in the text:

> To establish functional categories for all human genes
>
> To analyze genetic variations between humans, including the identification of single-nucleotide polymorphisms (SNPs)
>
> To map and sequence the genomes of several model organisms used in experimental genetics, including *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus* (the mouse)
>
> To develop new sequencing technologies, such as high-throughput computer-automated sequencers, to facilitate genome analysis
>
> To disseminate genome information, among both scientists and the general public

12. Copy number variations (CNVs) are duplications or deletions of large sections of repetitive DNA. Because many CNVs are not directly involved in production of a phenotype, they tend to be isolated from selection and show considerable variation in redundancy. Length variation in such repeats is unique among individuals (except for identical twins) and, with various detection methods, provides the basis for DNA fingerprinting. Single-nucleotide polymorphisms also occur frequently in the genome and can be used to distinguish individuals.

14. Whole genome sequencing (the strategy used for the HGP) provides sequencing results for entire genomes, including noncoding regions, whereas whole-exome sequencing provides sequence information only for exons. Since there are more disease-related variations in the exome than in other regions of the genome, WES is more likely to identify these mutations than is WGS. However, only WGS is able to identify mutations in regulatory regions that lead to disease.

16. A number of new subdisciplines of molecular biology will provide the infrastructure for major advances in our understanding of living systems. The following terms identify specific areas within that infrastructure:

proteomics—proteins in a cell or tissue
metabolomics—enzymatic pathways
glycomics—carbohydrates of a cell or tissue
toxicogenomics—toxic chemicals
metagenomics—environmental issues
pharmacogenomics—customized medicine
transcriptomics—expressed genes

Many other "-omics" are likely in the future.

18. Most microarrays, known also as gene chips, consist of a glass slide that is coated, using a robotic system, with single-stranded DNA molecules. Some microarrays are coated with single-stranded sequences of expressed sequenced tags or DNA sequences that are complementary to gene transcripts. A single microarray can have as many as 20,000 different spots of DNA (or as many as 1 million for exon-specific arrays), each containing a unique sequence. Researchers use microarrays to compare patterns of gene expression in tissues under different conditions or to compare gene-expression patterns in normal and diseased tissues. In addition, microarrays can be used to identify pathogens.

20. (a) Over 290,000 nonredundant peptides were identified from multiple organs, tissues, and cell types from clinically healthy individuals.
(b) Seven fetal tissues were used.
(c) A wide variety of searches can be performed.

22. (a) arm of human, wing of bird, fin of whale
(b) It happens that homologous proteins may have similar functions; however, there are many examples where different functions have evolved for homologous proteins.
(c) If proteins evolved by convergent evolution, they may have similar functions, but not share homology. Such proteins may have quite different amino acid sequences, but within general functional groupings and quite different DNA sequences. Code degeneracy may account for some of the DNA sequence variation.

24. In general, one would expect certain factors (such as heat or salt) to favor evolution to increase protein stability: distribution of ionic interactions on the surface, density of hydrophobic residues and interactions, and number of hydrogen and disulfide bonds. As seen from examining the codon table, a high GC ratio would favor the amino acids Ala, Gly, Pro, Arg, and Trp and minimize the use of Ile, Phe, Lys, Asn, and Tyr. How codon bias influences actual protein stability is not yet understood. Most genomic sequences change by relatively gradual responses to mild selection over long periods of time. They strongly resemble patterns of common descent; that is, they are conserved. Although the same can be said for organisms adapted to extreme environments, extraordinary physiological demands may dictate unexpected sequence bias.

26. According to the PANTHER database, 6 percent of human genes encode transcription factors; 3.3 percent encode cytoskeletal proteins; and 0.3 percent encode transmembrane receptor regulatory/adaptor proteins.

CHAPTER 22

Answers to Now Solve This

22.1 Antigens are usually quite large molecules, and in the process of digestion, they are sometimes broken down into smaller molecules, thus possibly becoming ineffective in stimulating the immune system. Some individuals are allergic to the food they eat, testifying to the fact that all antigens are not completely degraded or modified by digestion. In some cases, ingested antigens do indeed stimulate the immune system (e.g., oral polio vaccine) and provide a route for immunization. Localized (intestinal) immunity can sometimes be stimulated by oral introduction of antigens, and in some cases, this can offer immunity to ingested pathogens.

22.2 It will hybridize by base complementation to the normal DNA sequence.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of genomic applications that relate to agriculture, health and welfare, scientific exploration, and appreciation of the earth's flora and fauna, etc. In addition, areas of patent protection, personal privacy, and potential agricultural and environmental hazards should be addressed.

4. (a) Both the saline and column extracts of Lkt50 appear to be capable of inducing at least 50 percent neutralization of toxicity when injected into rabbits.
(b) First, the immunogen must be stably incorporated into the host plant hereditary material, and the host must express only that immunogen. During feeding, the immunogen must be transported across the intestinal wall unaltered or altered in such a way as to stimulate the desired immune response. There must be guarantees that potentially harmful by-products of transgenesis have not been produced. In other words, broad ecological and environmental issues must be addressed to prevent a transgenic plant from becoming an unintended vector for harm to the environment or any organisms feeding on the plant (directly or indirectly).

6. Since both mutations occur in the CF gene, children who possess both alleles will suffer from CF. Since both parents are heterozygotes, there is a 25 percent chance that any child will inherit both mutant alleles and develop CF.

8. Such widespread screening of newborns would allow the identification of a virtually infinite number of variables associated with the human genome that might be of scientific and personal interest. Some bases for certain disease states might be identified. However, disadvantages would be the likely stigmatizing of certain individuals and numerous issues of privacy invasion.

10. ASOs are short oligonucleotides that hybridize to their complementary strands in the genome. Since methylation does not interfere with base pairing, an ASO test would be unable to identify epigenetic DNA modifications.

12. Genome-wide association studies involve scanning the genomes of thousands of unrelated individuals with a particular disease and comparing them with the genomes of individuals who do not have the disease. GWAS attempt to identify genes that influence disease risk.

14. One can imagine that the existence of a synthetic human genome could usher in applications that would benefit humanity. For example, such technology might make it possible to easily "grow" synthetic transplant organs, eliminating incompatibility issues and waiting lists. However, there are a number of ethical concerns, which might outweigh the benefits.

16. In the case of haplo-insufficient mutations, gene therapy holds promise; however, in "gain-of-function" mutations, it is likely that the mutant gene's activity or product is compromised. RNAi or gene editing strategies may prove to be more effective. Addition of a normal gene will probably not help unless it can out-compete the mutant gene product.

18. Certainly, information provided to physicians and patients about genetic testing is a strong point in favor of wide distribution. It would probably be helpful for companies involved in genetic testing to participate by providing information peculiar to their operations. It would also be helpful if pooled statistical data were made available to the public in terms of frequencies of false positives and negatives, as well as population and/or geographic distributions. It would be necessary, however, that any individual results from tests would be held in strict confidence.

20. It is a highly personal decision to have one's genome sequenced, but in doing so one must be armed with information as to the expected variability of the genome and the possibility of false positives. A publically available genome might lead to employment bias, changes in personal relationships, and so on.

22. Services currently available from 23andMe include analysis of ancestry and health, with access to your raw data. Raw genomic information is difficult to interpret, and regulation of such companies, especially of the health-related information given, could be beneficial to protect consumers' interests.

24. (a) Since a gene is a product of the natural world, it does not conform to section 101 of U.S. patent laws, which govern patentable matter. Since the direct-to-consumer test for the *BRCA1* and *BRCA2* genes is original in its process or development, it should be patentable.

(b) Like the DTC test in part (a), Venter's "first-ever human-made life form" is original and patentable. While Venter's artificial genome may appear to be the same as the *BRCA1* and *BRCA2* genes, the former involves an original design process and the latter are works of nature: The genes themselves should not be patentable.

26. The ability to synthesize infective viruses *in vitro* effectively eliminates the possibility of the global eradication of pathogenic viruses. In addition, synthetic viruses remain pathogenic. Therefore, the possibility of their accidental—or deliberate—release is of great concern. Given that experiments on synthetic poliovirus were able to attenuate virulence, it is also possible that synthetic pathogens could be constructed to be more virulent—and perhaps also less responsive or completely unresponsive to current treatments and vaccines.

CHAPTER 23

Answers to Now Solve This

23.1 You may have identified several different mutations (multiple alleles) in some of the same genes. Furthermore, it is possible that your screen was more inclusive; that is, it identified more subtle alterations than the screen of others. Finally, your screen may have included some zygotic-effect mutations, which were dependent on the action of maternal-effect genes.

23.2 Because in *ftz/ftz* embryos, the engrailed product is absent and in *en/en* embryos, *ftz* expression is normal, one can conclude that the *ftz* gene product regulates *en*, either directly or indirectly. Because the *ftz* gene is expressed normally in *en/en* embryos, the product of the *engrailed* gene does not regulate expression of *ftz*.

Solutions to Problems and Discussion Questions

2. Your essay should include a description and examples of differential gene activity, homeotic genes, vulval development, and a typical signaling pathway such as Notch.

4. The fact that nuclei from almost any source remain transcriptionally and translationally active substantiates the fact that the genetic code, transcription, and translation are compatible throughout the animal and plant kingdoms. Because the egg represents an isolated, "closed" system that can be mechanically, environmentally, and to some extent biochemically manipulated, various conditions may develop that allow one to study facets of gene regulation. For instance, the influence of transcriptional enhancers and suppressors may be studied along with factors that impact translational and posttranslational processes. Combinations of injected nuclei may reveal nuclear—nuclear interactions, which could not normally be studied by other methods.

6. (a) Genes that control early development are often dependent on the deposition of their products (mRNA, transcription factors, various structural proteins, etc.) in the egg by the mother. When observable, these are maternal-effect genes.

(b) They are made in the early oocyte or nurse cells during oogenesis and deposited in gradients throughout the egg.

(c) Such maternal-effect genes control early developmental events such as defining anterior—posterior polarity. Such products are placed in eggs during oogenesis and are activated immediately after fertilization.

(d) A variety of phenotypes are possible, including embryonic lethality, and they are often revealed in the offspring of females. Maternal effects reveal the genotype of the mother: for example, females homozygous for deleterious recessive mutations of maternal-effect genes are sterile.

8. The three main classes of zygotic genes are (1) *gap* genes, which specify adjacent segments; (2) *pair-rule* genes, which specify every other segment and a part of each segment; and (3) *segment polarity* genes, which specify homologous parts of each segment.

10. There are several somewhat indirect methods for determining the transcriptional activity of a given gene in different cell types. First, if protein products of a given gene are present in different cell types, it can be assumed that the responsible gene is being transcribed. Second, if one is able to actually observe, microscopically, gene activity, as is the case in some specialized chromosomes (polytene chromosomes), gene activity can be inferred by the presence of localized chromosomal puffs. Third, a more direct and common practice to assess transcription of particular genes is to use labeled probes. If a labeled probe can be obtained that contains base sequences that are complementary to the transcribed RNA, then such probes will hybridize to that RNA if present in different tissues. This technique is called *in situ* hybridization and is a powerful tool.

12. A dominant gain-of-function mutation is one that changes the specificity or expression pattern of a gene or gene product. The "gain-of-function" *Antp* mutation causes the wild-type *Antennapedia* gene to be expressed ectopically in the eye-antenna disc and mutant flies have legs on the head in place of antenna.

14. Because of the regulatory nature of *homeotic* genes in fundamental cellular activities of determination and differentiation, it would be difficult to ignore their possible impact on oncogenesis. Homeotic genes encode DNA binding domains, which influence gene expression; any factor that influences gene expression may, under some circumstances, influence cell-cycle control.

However attractive this model, there have been no homeotic transformations noted in mammary glands, so the typical expression of mutant homeotic genes in insects is not revealed in mammary tissue, according to Lewis (2000). A substantial number of experiments will be needed to establish a functional link between homeotic gene mutation and cancer induction. Mutagenesis and transgenesis experiments are likely to be most productive in establishing a cause-and-effect relationship.

16. First, one should know what types of mRNAs are being translated with these ribosomes. Second, it would be interesting to know whether inhibitors of mitochondrial-ribosomal translation would interfere with germ-cell formation.

18. Given the information in the problem, it is likely that this gene normally controls the expression of *BX-C* genes in all body segments. The wild-type product of *esc* stored in the egg may be required to interpret the information correctly stored in the egg.

20. Three classes of flower homeotic genes specify various floral organs. Class *A* group gives rise to petals. *B* and *C* genes control stamen formation, and *C* genes regulate carpels.

(a) *n300* most likely functions earlier than *let-23*.

(b) The double mutant should be vulvaless because both genes block vulva formation.

22. (a) *n300* most likely functions earlier than let-23.

(b) The double mutant should be vulvaless because both genes block vulva formation.

24. (a,b) A number of studies indicate that genes in *Drosophila* have evolutionary counterparts (orthologs) in other organisms, including humans. A number of similar genes influence eye development in both insects and vertebrates. Genes that produce eyes are part of a complex network of at least seven genes that constitute the master regulators of eye development. Each gene functions in coordination with others in a conserved network that is used by broad evolutionary groups. Such genes, descended from common ancestral genes that have the same function in different species, are called orthologs.

CHAPTER 24

Answers to Now Solve This

24.1 Several approaches are used to combat CML. One includes the use of a tyrosine kinase inhibitor that binds competitively to the ATP binding site of ABL kinase, thereby inhibiting phosphorylation of BCR-ABL and preventing the activation of additional signaling pathways. In addition, real-time quantitative reverse transcription-polymerase chain reaction (Q-RT-PCR) allows one to monitor the drug responses of cell populations in patients so that less toxic and more effective treatments are possible. Being able to distinguish leukemic cells from healthy cells allows one not only to target therapy to specific cell populations, but also to quantify responses to therapy. Because such cells produce a hybrid protein, it may be possible to develop a therapy, perhaps an immunotherapy, based on the uniqueness of the BCR-ABL protein.

24.2 *TP53* is a tumor-suppressor gene that protects cells from multiplying with damaged DNA. It is present in its mutant state in more than 50 percent of all tumors. Since the immediate control of a critical and universal cell-cycle checkpoint is mediated by *TP53*, mutation will influence a wide range of cell types. *TP53*'s action is not limited to specific cell types.

24.3 Cancer is a complex alteration in normal cell-cycle controls. Even if a major "cancer-causing" gene is transmitted, other genes, often new mutations, are usually necessary in order to drive a cell toward tumor formation. Full expression of the cancer phenotype is likely to be the result of an interplay among a variety of genes and will therefore show variable penetrance and expressivity.

24.4 Obviously, it is less expensive, both in terms of human suffering and money, to seek preventive measures for as many diseases as possible. Therefore, it is extremely important that we increase efforts to educate and protect the human population from as many hazardous environmental agents as possible. However, having gained some understanding of the mechanisms of disease, in this case cancer, it must also be stated that no matter what preventive measures are taken it will be impossible to completely eliminate disease from the human population. A balanced, multipronged approach seems appropriate.

Solutions to Problems and Discussion Questions

2. Your essay should describe the general influence of genetics in cancer. Since epigenetic factors alter gene output, it is likely that such factors could cause cancer.

4. Kinases regulate other proteins by adding phosphate groups. Cyclins bind to the kinases, switching them on and off. CDK4 (cyclin-dependent kinase) binds to cyclin D, moving cells from G1 to S. At the G2/mitosis border, CDK1 combines with another cyclin (cyclin B). The resulting phosphorylation brings about a series of changes in the nuclear membrane via caldesmon, cytoskeleton, and histone H1.

6. To say that a particular trait is inherited conveys the assumption that when a particular genetic circumstance is present, it will be revealed in the phenotype. When one discusses an inherited predisposition, one usually refers to situations where a particular phenotype is expressed in families in some consistent pattern. However, the phenotype may not always be expressed or may manifest itself in different ways.

8. Apoptosis, or programmed cell death, is a genetically controlled process that leads to death of a cell. It is a natural process involved in morphogenesis and a protective mechanism against cancer formation. During apoptosis, nuclear DNA becomes fragmented, cellular structures are disrupted, and the cells are dissolved. Caspases are involved in the initiation and progress of apoptosis.

10. Normally, the p53 protein is bound by the MDM2 protein, which marks p53 for degradation and prevents its activation. Upon DNA damage or cellular stress, MDM2 dissociates from p53, which both stabilizes the protein and activates it. Once activated, p53 serves as a transcription factor, upregulating genes that lead to cell-cycle arrest (*p21*) and apoptosis (*BAX*). Individuals with two copies of a tumor suppressor gene would need to experience separate mutations in both copies to develop cancer, whereas individuals with only one functional copy (plus one mutant copy) would only need a single mutation. Therefore, it make sense that those who inherit one mutant copy of a recessive tumor-suppressor gene will have the higher risk of developing cancer.

12. If DNA replication or repair of any DNA damage has not been completed, the cell cycle is arrested at the G2/M checkpoint (mediated by activated p53 protein) until these processes are complete. This allows time for DNA damage to be repaired during the S phase. In addition, p53 upregulates genes, such as *GADD45*, that mediate DNA repair.

14. Mutations that produce oncogenes alter gene expression either directly or indirectly and act in a dominant capacity. Proto-oncogenes are those that normally function to promote or maintain cell division. In the mutant state (oncogenes), they induce or maintain uncontrolled cell division; that is, there is a gain of function. Generally, this gain of function takes the form of increased or abnormally continuous gene output. On the other hand, loss of function is generally attributed to mutations in tumor-suppressor genes, which function to halt passage through the cell cycle. When such genes are mutant, they have lost their capacity to halt the cell cycle. Such mutations are generally recessive.

16. To encourage infected cells to undergo growth and division, viruses often encode genes that stimulate growth and division. Many viruses either inactivate tumor-suppressor genes of the host or bring in genes that stimulate cell growth and division. By inactivating tumor-suppressor genes, the normal braking mechanism of the cell cycle is destroyed.

18. Normal cells are often capable of withstanding mutational assault because they have checkpoints and DNA repair mechanisms in place. When such mechanisms fail, cancer may be a result. Through mutation, such protective mechanisms are compromised in cancer cells, and as a result they show higher than normal rates of mutation, chromosomal abnormalities, and genomic instability.

20. Epigenetic effects can be caused by DNA methylation and/or histone modifications, including acetylation and/or phosphorylation. As such, they can activate or silence whole chromosomes (X chromosome, for example) or certain chromosomal regions; they can be responsible for parental imprinting; and they can influence gene activity in heterochromatin. Patterns of nucleotide demethylation and hypermethylation are often different when cancer cells are compared to normal cells, as are histone modifications.

22. No, she will still have the general population risk of about 10 percent. In addition, it is possible that genetic tests will not detect all breast cancer mutations.

24. Proteases in general and serine proteases, specifically, are considered tumor-promoting agents because they degrade proteins, especially those in the extracellular matrix. When such proteolysis occurs, cellular invasion and metastasis are encouraged. Consistent with this observation are numerous observations that metastatic tumor cells are associated with higher than normal amounts of protease expression. Inhibitors of serine proteases are often tested for their anticancer efficacy.

26. As with many forms of cancer, a single gene alteration is not the only requirement. The authors (Bose et al.) state "but only infrequently do the cells acquire the additional changes necessary to produce leukemia in humans." It is possible that the cells that produce these transcripts are extremely low in number and are unable to establish themselves as key stem cells for clonal expansion. It is also possible that the transcripts produced do not result in a functional fusion protein.

28. (a) The double-stranded DNA sequence for a glutamine codon is either $5' - \text{CAG} - 3'/3' - \text{CTG} - 5'$ or $5' - \text{CAA} - 3'/3' - \text{TTG} - 5'$. Therefore, the mRNA triplet for Gln is CAG(A). The mRNA triplet that specifies a stop is one of three: UAA, UAG, or UGA. In the double-stranded DNA, the C-G base pair at the 5′ end of the Gln codon would need to change to a T-A base pair. This could occur either by a nucleotide change from C to T on the coding strand or by a change from G to A on the template strand. In either case, if replication occurs, one double-stranded daughter DNA molecule will contain a TAG(A)/ATC(T) sequence. When transcribed, this would produce a UAG(A) triplet in the transcript, which would signal a translational stop.

(b) It is likely to be a tumor-suppressor gene because loss of function causes predisposition to cancer.

(c) Some women may carry genes (perhaps mutant) whose products provide functions that are the same as that of the *BRCA1* gene product. Some women may have immune systems that recognize and destroy precancerous cells, or they may have mutations in breast signal transduction genes so that suppression of cell division occurs in the absence of *BRCA1*.

30. Three circumstances described in the article include transfer of cancerous cells during an organ transplant, transfer of cancer from mother to fetus during pregnancy, and transfer of pathogens that are linked to cancer formation. Refer to the article for a more detailed description of these routes to cancer formation. One condition shared by these routes is the weakened immune system of the recipient. Refer to the article for a more detailed explanation.

CHAPTER 25

Answers to Now Solve This

25.1 (a) Since 1/256 of the F_2 plants are 20 cm and 1/256 are 40 cm, there must be four gene pairs involved in determining flower size.

(b) The backcross is: *AaBbCcDc* × *AABBCCDD*. The frequency distribution in the backcross would be:

1/16 = 40 cm
4/16 = 37.5 cm
6/16 = 35 cm
4/16 = 32.5 cm
1/16 = 30 cm

25.2 (a) Taking the sum of the values and dividing by the number in the sample gives the following means:

mean sheep fiber length = 7.7 cm
mean fleece weight = 6.4 kg

The variance for each is:

variance sheep fiber length = 6.097
variance fleece weight = 3.12

The standard deviation is the square root of the variance:

sheep fiber length = 2.469
fleece weight = 1.766

(b,c) The covariance for the two traits is 29.76/7, or 4.25, while the correlation coefficient is + 0.974.

(d) There is a very high correlation between fleece weight and fiber length, and it is likely that this correlation is not by chance. Even though correlation does not mean cause and effect, it would seem logical that as you increased fiber length, you would also increase fleece weight. It is probably safe to say that the increase in fleece weight is directly related to an increase in fiber length.

25.3 Compare the expression of traits in monozygotic and dizygotic twins. A higher concordance value for monozygotic twins indicates a significant genetic component for a given trait. Notice that for traits including blood type, eye color, and mental retardation, there is a fairly significant difference between the MZ and DZ groups. However, for measles and handedness, the difference is not as significant, indicating a greater role of the environment. Hair color has a significant genetic component, as do idiopathic epilepsy, schizophrenia, diabetes, allergies, cleft lip, and club foot. The genetic component of mammary cancer is present but minimal according to these data.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of various ratios typical of Mendelian genetics as compared with the more blending, continuously varying expressions of neo-Mendelian modes of inheritance. It should contrast discontinuous inheritance and continuous patterns.

4. (a) Determine the number of genes either from the ratio of F_2 individuals expressing either extreme phenotype or from the number of phenotypic categories. Either method indicates there are two alleles at each locus, for a total of four alleles.

(b,c) Because the description of red, medium red, and so on gives us no indication of a *quantity* of color in any form of units, we would not be able to actually quantify a unit amount for each change in color. We can say that each gene (additive allele) provides an equal unit amount to the phenotype, and the colors differ from each other in multiples of that unit amount. The number of additive alleles needed to produce each phenotype is as follows:

1/16 = dark red	= *AABB*	4 additive alleles
4/16 = medium-dark red	= 2*AABb* 2*AaBB*	3 additive alleles
6/16 = medium red	= *AAbb* 4*AaBb* *aaBB*	2 additive alleles
4/16 = light red	= 2*aaBb* 2*Aabb*	1 additive allele
1/16 = white	= *aabb*	0 additive alleles

(d) F_1 = all light red
F_2 = 1/4medium red
2/4 light red
1/4 white

6. (a,b) When four gene pairs act additively, the proportion of either extreme phenotype to the total number of offspring is 1/256. The extreme types in this problem are the 12-cm and 36-cm plants. This observation suggests that there are four gene pairs involved.

(c) If there are four gene pairs, there are nine ($2n + 1$) phenotypic categories and eight increments between these categories. Since there is a difference of 24 cm between the extremes, 24 cm/8 = 3 cm for each increment (each of the additive alleles).

(d) Because the parents are inbred, it is expected that they are fully homozygous. An example:

AABBccdd × *aabbCCDD*

(e) Since the *aabbccdd* genotype gives a height of 12 cm and each uppercase allele adds 3 cm to the height, there are many possibilities for an 18-cm plant:

AAbbccdd,
AaBbccdd,
aaBbCcdd, etc.

Any plant with seven uppercase letters will be 33 cm tall:

AABBCCDd,
AABBCcDD,
AABbCCDD, for example.

8. For height, notice that average differences between MZ twins reared together (1.7 cm) and MZ twins reared apart (1.8 cm) are similar (meaning little environmental influence) and considerably less than differences of DZ twins (4.4 cm) or sibs (4.5) reared together. These data indicate that genetics plays a major role in determining height.

However, for weight, notice that MZ twins reared together have a much smaller (1.9 kg) difference than MZ twins reared apart, indicating that the environment has a considerable impact on weight. By comparing the weight differences of MZ twins reared apart with DZ twins and sibs reared together, one can conclude that the environment has almost as much an influence on weight as genetics.

For ridge count, the differences between MZ twins reared together and those reared apart are small. For the data in the table, it would appear that ridge count and height have the highest heritability values.

10. Many traits, especially those we view as quantitative, are likely to be determined by a polygenic mode with possible environmental influences. The following are some common examples: height, general body structure, skin color, and perhaps most common behavioral traits, including intelligence.

12. (a) *For back fat:*

Broad-sense heritability $= H^2 = 12.2/30.6 = 0.398$
Narrow-sense heritability $= h^2 = 8.44/30.6 = 0.276$

For body length:

Broad-sense heritability $= H^2 = 26.4/52.4 = 0.504$
Narrow-sense heritability $= h^2 = 11.7/52.4 = 0.223$

(b) Selection for back fat would produce more response.

14. (a) For vitamin A: $h_{A^2} = 0.097$
For cholesterol: $h_{A^2} = 0.223$

(b) Cholesterol content should be influenced to a greater extent by selection.

16. $h^2 = (7.5 - 8.5/6.0 - 8.5) = 0.4$(realized heritability)

18. $h^2 = 0.3 = (M2 - 60/80 - 60)\ M2 = 66$g

20. (a,b) In many instances, a trait may be clustered in families, yet traditional mapping procedures may not be applicable because the trait might be influenced by a number of genes. In general, researchers look for associations to particular DNA sequences (molecular markers). *Cosegregation* is said to occur when the phenotypic trait and the molecular marker are genetically linked. When the trait cosegregates with a particular marker and it statistically associates with that trait above chance, a likely QTL has been identified. Markers such as RFLPs, SNPs, and microsatellites are often used because they are chromosome-specific, highly variable, relatively easy to assess, and present in all individuals.

22. The solution to these types of problems rests on determining the ratio of individuals expressing the extreme phenotype to the total number of individuals. In this case, 8:2028 is equal to 1:253, which is close to 1:256. If there are three gene pairs, the ratio is 1:64; four gene pairs, 1:256; or five gene pairs, 1:1024. Therefore, these data indicate that four gene pairs influence size in these guinea pigs.

24. (a,b) Because there are nine phenotypic classes in the F_2, four gene pairs must be involved. The genotypes of the parents could be symbolized as *AABBCCDD* × *aabbccdd*, and the F_1 as *AaBbCcDd*. Note that there are additional possible parental genotypes that will yield F_1 individuals that are heterozygous at all four loci.

26. $6 \times 5 \times 4 \times 3 \times 2 \times 1/(2 \times 1)(4 \times 3 \times 2 \times 1) = 15$ of a total of 64.

28. (a) The average response to selection (in mm) would be the sum of the differences between the control and offspring, divided by 3: $(2.17 + 3.79 + 4.06)/3 = 3.34$ mm.

(b) One computes realized heritability by the following formula:

$$h^2 = \frac{M2 - M}{M1 - M}$$

where M represents the mean size (control), $M1$ represents the selected parents, and $M2$ represents the size in offspring.

For 1997:

$$h^2 = (32.21 - 30.04)/(34.13 - 30.04) = 0.53$$

For 1998:

$$h^2 = (31.90 - 28.11)/(31.98 - 28.11) = 0.98$$

For 1999:

$$h^2 = (33.74 - 29.68)/(31.81 - 29.68) = 1.91$$

The overall realized heritability would be the average of all three heritability values, or 1.14.

(c) A key factor in determining response to selection is the genetic variability available to respond to selection. The greater the genetic variability, the greater the response. Another key factor in rapid response to selection often relates to the number of loci involved. If there are few loci, each with large phenotypic effects controlling a trait, response to selection is usually high. Finally, if flower size is not genetically correlated with other floral traits, size alone may not be subject to strong stabilizing selection, which would reduce genetic variation. Therefore, whereas other floral traits may show low response to selection, size alone may be more responsive.

(d) With high heritability, one expects a high genetic contribution to phenotypic variation. Genetic variability that provides rapid adjustments to changing environments is an evolutionary advantage. Therefore, in general, one would expect that high heritability would contribute to high evolutionary potential.

CHAPTER 26

Answers to Now Solve This

26.1 Because the alleles follow a dominant/recessive mode, one can use the equation $\sqrt{q^2}$ to calculate *q*, on which all other aspects of the answer depend. The frequency of *aa* types is determined by dividing the number of nontasters (37) by the total number of individuals (125).

$$q^2 = 37/125 = 0.296$$
$$q = 0.54$$
$$p = 1 - q$$
$$p = 0.456$$

The frequencies of the genotypes are determined by applying the formula $p^2 + 2pq + q^2$ as follows:

$$\begin{aligned}\text{Frequency of } AA &= p^2\\ &= (0.456)^2\\ &= 0.208 \text{ or } 20.8\%\\ \text{Frequency of } Aa &= 2pq\\ &= 2(0.456)(0.544)\\ &= 0.496 \text{ or } 49.6\%\\ \text{Frequency of } aa &= q^2\\ &= (0.544)^2\\ &= 0.296 \text{ or } 29.6\%\end{aligned}$$

When completing such a set of calculations, it is a good practice to add the final percentages to be certain that they total 100 percent. (Note that the calculation requires the assumption that this population is in Hardy–Weinberg equilibrium with respect to the gene for PTC tasting.)

26.2 (a) For the CCR5 analysis, first determine p and q. Knowing the frequencies of all the genotypes, one can calculate p as the sum of 0.6 and $(0.351/2) = 0.7755$; q will be the sum of 0.049 and $(0.351/2) = 0.2245$.
The equilibrium values will be as follows:

Frequency of $1/1 = p^2 = (0.7755)^2 = 0.6014$ or 60.14%
Frequency of $1/\Delta\ 32 = 2pq = 2(0.7755)(0.2245)$ $= 0.3482$ or 34.82%
Frequency of $\Delta 32/\Delta\ 32 = q^2 = (0.2245)^2 = 0.0504$ or 5.04%

Comparing these equilibrium values with the observed values strongly suggests that the observed values are drawn from a population in Hardy–Weinberg equilibrium.

(b) For the sickle-cell analysis, first determine p and q. Since the frequencies of all the genotypes are known, one can calculate p as the sum of 0.756 and $(0.242/2) = 0.877$; q will be $(1 - 0.877)$ or 0.123. [Note: You can also calculate q as the sum of 0.002 and $(0.242/2) = 0.123$.]
The equilibrium values will be as follows:

Frequency of $SS = p^2 = (0.877)^2 = 0.7691$ or 76.91%
Frequency of $Ss = 2pq = 2(0.877)(0.123) = 0.2157$ or 21.57%
Frequency of $ss = q^2 = (0.123)^2 = 0.0151$ or 1.51%

Comparing these equilibrium values with the observed values suggests that the observed values may be drawn from a population that is not in equilibrium. Notice that there are more heterozygotes and fewer homozygotes (especially the *ss* types) than predicted in the population. Because data are given in percentages, χ^2 values cannot be computed.

26.3 The recessive allele a is present in the homozygous state (q^2) at a frequency of 0.0001.

(a) q is 0.01
(b) $p = 1 - q$ or 0.99
(c) $2pq = 2(0.01)(0.99) = 0.0198$ (or about 1/50)
(d) $2pq \times 2pq = 0.0198 \times 0.0198 = 0.000392$ (or about 1/255)

26.4 The overall probability of the couple producing a CF child is $98/2500 \times 2/3 \times 1/4 = 0.00653$, or about 1/153.

Solutions to Problems and Discussion Questions

2. Your essay should include a discussion of the original sources of variation coming from mutation and that migration can cause gene frequencies to change in a population if the immigrants have different gene frequencies compared to the host population. You should also describe selection as resulting from the biased passage of gametes and offspring to the next generation.

4. The classification of organisms into different species is based on evidence (morphological, genetic, ecological, etc.) that they are reproductively isolated. That is, there must be evidence that gene flow does not occur among the groups being called different species. Classifications above the species level (genus, family, etc.) are not based on such empirical data. Indeed, classification above the species level is somewhat arbitrary and based on traditions that extend far beyond DNA sequence information. In addition, recall that DNA sequence divergence is not always directly proportional to morphological, behavioral, or ecological divergence. Although the genus classifications provided in this problem seem to be invalid, other factors, well beyond simple DNA sequence comparison, must be considered in classification practices. As more information is gained on the meaning of DNA sequence differences in comparison to morphological factors, many phylogenetic relationships will be reconsidered, and it is possible that adjustments will be needed in some classification schemes.

6. For each of these values, one merely takes the square root to determine q, then computes p, and then "plugs" the values into the $2pq$ expression.

(a) $q = 0.08$ $\quad 2pq = 2(0.92)(0.08) = 0.1472$ or 14.72%

(b) $q = 0.009$ $\quad 2pq = 2(0.991)(0.009) = 0.01784$ or 1.78%

(c) $q = 0.3$ $\quad 2pq = 2(0.7)(0.3) = 0.42$ or 42%

(d) $q = 0.1$ $\quad 2pq = 2(0.9)(0.1) = 0.18$ or 18%

(e) $q = 0.316$ $\quad 2pq = 2(0.684)(0.316) = 0.4323$ or 43.23%

(Depending how one rounds off the decimals, slightly different answers will occur.)

8. Assuming that the population is in Hardy–Weinberg equilibrium, if one has the frequency of individuals with the dominant phenotype, the remainder has the recessive phenotype (q^2). With q^2, one can calculate q, and from this value one can arrive at p. Applying the expression $p^2 + 2pq + q^2$ will allow a solution to the question.

10. The following formula calculates the frequency of an allele in the next generation for any selection scenario, given the frequencies of a and A in this generation and the fitness of all three genotypes; $q_{g+1} = [w_{Aa}p_gq_g + w_{aa}q_g{}^2]/[w_{AA}p_g{}^2 + w_{Aa}2p_gq_g + w_{aa}q_g{}^2]$, where q_{g+1} is the frequency of the a allele in the next generation, q_g is the frequency of the a allele in this generation, p_g is the frequency of the A allele in this generation, and each "w" represents the fitness of its respective genotype.

(a) $q_{g+1} = [0.9(0.7)(0.3) + 0.8(0.3)2]/[1(0.7)_2 + 0.9(2)(0.7)(0.3) + 0.8(0.3)_2]$

$q_{g+1} = 0.278\ p_{g+1} = 0.722$

(b) $q_{g+1} = 0.289\ p_{g+1} = 0.711$
(c) $q_{g+1} = 0.298\ p_{g+1} = 0.702$
(d) $q_{g+1} = 0.319\ p_{g+1} = 0.681$

12. Because a dominant lethal gene is highly selected against, it is unlikely that it will exist at too high a frequency, if at all. However, if the gene shows incomplete penetrance or late age of onset (after reproductive age), it may remain in a population.

14. The frequency of an allele is determined by a number of factors, including the fitness it confers, mutation rate, and input from migration. There is no tendency for a gene to reach any artificial frequency such as 0.5. In fact, you have seen that rare alleles tend to remain rare even when they are dominant—unless there is very strong selection for the allele. The distribution of a gene among individuals is determined by mating (population size, inbreeding, etc.) and environmental factors (selection, etc.). A population is in Hardy–Weinberg equilibrium when the distribution of genotypes occurs at or around the $p^2 + 2pq + q^2 = 1$ expression. Equilibrium does not mean 25 percent *AA*, 50 percent *Aa*, and 25 percent *aa*. This confusion often stems from the 1:2:1 (or 3:1) ratio seen in Mendelian crosses.

16. The frequency of mutation is 2/100,000, or 2×10^{-5}.

18. The presence of selectively neutral alleles and genetic adaptations to varied environments contribute significantly to genetic variation in natural populations.

20. The frequency of the *b* allele after one generation of corn borers fed on Bt corn would be computed as follows:

$$q_{g+1} = {}^{[w_{Bb}p_g q_g + w_{bb}q_g{}^2]}\!\Big/\!{}_{[w_{BB}p_g{}^2 + w_{Bb}2p_g q_g + w_{bb}q_g{}^2]}$$

$$\begin{aligned} q_{g+1} &= [1(0.02)(0.98) + (0.1)(0.98)^2]/[1(0.02)^2 \\ &\quad + 1(2)(0.02)(0.98) + (0.1)(0.98)^2] \\ q_{g+1} &= 0.852 \text{ and } p_{g+1} = 0.148 \end{aligned}$$

22. Use the formula $p_i2 = (1 - m)p_i + mp_m$.

$$\begin{aligned} p_i' &= (1 - 0.17)(0.75) + (0.167)(1.0) \\ p_i' &= 0.792 \text{ and } q_i' = 0.208 \end{aligned}$$

Note that the value of 0.167 comes from the fact that 10 sheep are introduced into an existing population of 50, so 10/60 (or 16.7 percent) of the parents of the next generation are migrants. Also note that there are no *cc* alleles in the introduced population, so $p_m = 1.0$.

24. Somatic gene therapy, like any therapy, allows some individuals to live more normal lives than those not receiving therapy. As such, the ability of these individuals to contribute to the gene pool increases the likelihood that less-fit alleles will enter and be maintained in the gene pool. This is a normal consequence of therapy, genetic or not, and in the face of disease control and prevention, societies have generally accepted this consequence. Germ-line therapy could, if successful, lead to limited, isolated, and infrequent removal of an allele from a gene lineage. However, given the present state of the science, its impact on the course of human evolution will be diluted and negated by a host of other factors that afflict humankind.

26. Reproductive isolating mechanisms are grouped into prezygotic and postzygotic. Prezygotic mechanisms are more efficient because they occur before resources are expended in the processes of mating.

28. In small populations, large fluctuations in gene frequency occur because random gametic sampling may not include all the genetic variation in the parents. The same phenomenon occurs in molecular populations. Two factors can cause the extinction of a particular mutation in small populations. First, sampling error may allow fixation of one form and the elimination of others. If an advantageous mutation occurs, it must be included in the next replicative round in order to be maintained in subsequent generations. If the founding population is small, it is possible that the advantageous mutation might not be represented. Second, although the previous statements also hold for deleterious mutations, if a deleterious mutation becomes fixed, it can lead to extinction of that population.

30. With one exception (the final amino acid in the sequences), all the amino acid substitutions (Ala – Gly, Val – Leu, Asp – Asn, Met – Leu) require only one nucleotide change. The last change from Pro (CC–) – Lys (AAA,G) requires two changes (the minimal mutational distance).

SPECIAL TOPIC 1

Review Question Answers

2. Bacteria that became resistant to a given phage strain acquired new spacer sequences in their CRISPR loci that matched portions of the phage genome. Deletion or mutation of these spacers is correlated with renewed sensitivity to the phage, and experimental insertion of these sequences into the CRISPR loci of sensitive bacteria made them resistant.

4. The type II CRISPR-Cas9 system of *S. pyogenes* was selected due to its simplicity. Only a few components are needed, chief among them being the Cas9 nuclease, which plays a role in all three steps of CRISPR-Cas-mediated defense.

6. A single guide RNA (sgRNA) is an engineered RNA molecule that takes the place of the crRNA/tracrRNA duplex. The sgRNA contains 20 nucleotides of a crRNA sequence linked to the minimal functional sequence of tracrRNA.

8. One example discussed in the text is the precise modification of the pig *CD163* gene to remove the domain of the receptor that interacts with PRRSV. This makes the pigs resistant to infection without compromising the beneficial immune functions of the CD163 receptor.

Discussion Question Answers

2. A specific 20-nucleotide sequence occurs once every 4^{20} or 1.1×10^{12} base pairs and a PAM (52-NGG-32) occurs once every 16 base pairs in a random sequence. Both will occur once every 1.76×10^{13} base pairs. Given that the haploid human genome is 3.2×10^9 bp, finding both by pure chance alone would be highly improbable.

4. Both systems rely on complementary base pairing to recruit a protein to target sequences, and in both, the guide molecule is a short RNA molecule that is derived from a larger transcript. The systems differ in their targets (mRNA versus genomic DNA). Also, while the miRNA/RISC system is limited to gene silencing, editing by CRISPR-Cas can result in silencing, removal of a single protein domain, or the substitution of a functional gene for a mutant gene.

6. Both cystic fibrosis and hemophilia would be good candidates for gene editing by CRISPR-Cas, since each is caused by a recessive mutation of a single gene. Down syndrome would be a poor candidate for treatment using CRISPR-Cas because the condition is caused by an extra copy of an entire chromosome (trisomy 21) or by translocation of a critical portion of chromosome 21. Brain cancer is also a poor candidate, since its cause is not known.

SPECIAL TOPIC 2

Review Question Answers

2. With the development of the polymerase chain reaction, trace samples of DNA can be used, commonly in forensic applications. STRs are like VNTRs, but the repeat portion is shorter, between two and nine base pairs, repeated from

7 to 40 times. A core set of STR loci, about 20, is most often used in forensic applications.

4. Since males typically contain a Y chromosome (exceptions include transgender and mosaic individuals), gender separation of a mixed tissue sample is easily achieved by Y chromosome profiling. In addition, STR profiling is possible for over 200 loci; however, because of the relative stability of DNA in the Y chromosome, it is difficult to differentiate between DNA from fathers and sons or male siblings.

6. Like Y chromosome DNA, mtDNA is very stable because it undergoes very little, if any, recombination. Since there is a high copy number of mitochondria in cells, it is especially useful in situations where samples are small, old, or degraded, which is often the case in catastrophes.

8. The Combined DNA Index System (CODIS) is a collection of DNA databases and analytical tools of both state and federal governments, maintained by the FBI. As of 2016, it contained more than 16 million DNA profiles. DNA profiles are collected from convicted offenders, forensic investigations, and in some states, those suspected of crimes as well as from unidentified human remains and missing persons (in cases where DNA is available).

10. Although useful, DNA profiling has a number of limitations. Many criminal cases do not have DNA evidence, or the DNA evidence they have is not useful. For those cases for which DNA *is* useful, samples can remain unprocessed. Human error—or deliberate human interference—can introduce errors into DNA evidence.

Discussion Question Answers

2. To gain information as to laws and regulations in various states, one could navigate to "Welcome to the DNA Laws Database" within the National Conference of State Legislatures Web site. There, one can select a particular state for its laws and regulations regarding DNA collection and profiling. In general, one will see that most states contain descriptions of the following topics:

(a) various DNA databases used
(b) methods of DNA collection
(c) postconviction DNA collection of felons
(d) oversight and advisory committees
(e) convicted offender statutes

4. Somatic mosaicism and chimerism involve a mixture of cell types, the origin of which may involve a variety of embryonic events, some of which are understood. Since a single individual may contain a mixed population of cells, a DNA sample taken from one tissue site may not match a DNA sample taken from another site. This can lead to a conflicted set of results when it comes to matching a DNA sample to a sample of DNA from a crime scene. Taking DNA samples from various sites on an individual may be useful in mitigating such confusion. In addition, in STR DNA profiling, mosaicism may present itself at the electrophoresis/analysis stage by additional peaks or peak height imbalances.

SPECIAL TOPIC 3

Review Question Answers

2. Herceptin is used in the treatment of breast cancer that targets the epidermal growth factor receptor 2 (*HER-2*) gene located on chromosome 17. Overexpression of this gene occurs in about 25 percent of invasive breast cancer cases. Herceptin is a monoclonal antibody that binds specifically to inhibit the HER-2 receptor. Because Herceptin acts only on cancer cells that have amplified *HER-2* genes, it is important to know the HER-2 status of each tumor. Molecular tests that are conducted include immunohistochemistry, which detects HER-2 receptors on tumor cells, and fluorescence *in situ* hybridization, which assesses the number of *HER-2* genes.

4. Indirect repression of T cells by cancer cells occurs when abnormal expression of MHC molecules allows cancer cells to avoid recognition by antigen-presenting cells, which prevents activation of T cells. Direct repression of T-cell activity occurs when cancer cells synthesize molecules that bind to and repress T cells. Another strategy is the presence of tumor-associated T regulatory cells (T regs) or tumor-infiltrating cells (macrophages and monocytes), all of which suppress T-cell activities.

6. Chimeric antigen receptors are genetically engineered proteins that are expressed on T cells and allow them to recognize tumor cells without activation by antigen-presenting cells. The chimeric proteins contain five key domains: a signal peptide, which directs the protein to the cell surface; the variable regions, which recognize and bind a specific antigen; a linker region, which allows the variable domains to position themselves correctly; a transmembrane domain; and the intracellular signaling domain. The protein is made by cloning DNA fragments encoding each domain into a single molecule and introducing this construct into naïve T cells of the patient.

8. There are 44,698 tests described on this Web site, most of which are used diagnostically. One example of single-gene testing occurs for patients with alpha-ketoglutarate dehydrogenase deficiency. The test is conducted for the *oxoglutarate dehydrogenase* gene, which encodes one subunit of the 2-oxoglutarate dehydrogenase complex. (Note that another name for 2-oxoglutarate is alpha-ketoglutarate.) The method employed is Sanger sequencing of both strands, looking for gene variants.

Discussion Question Answers

2. There are a number of technical challenges to be overcome. Current methodologies need to be faster, more accurate, and less expensive. In addition, coordinating the storage of and access to the vast amounts of data that will be generated will be daunting. Once the data are generated, scientists and health professionals will need to ensure that all relevant information is linked. In addition, nonmedical or genetic information, such as the effects of environment, life-style, and epigenetic contributions, will need to be included.

4. At present, genetic discrimination does exist; however, recent developments in health care laws seek to minimize such discrimination by medical insurance companies. It remains to be seen whether genetic discrimination in the workplace continues.

SPECIAL TOPIC 4

Review Question Answers

2. Genetic engineering allows genetic material to be transferred within and between species and to alter expression levels of genes. A transgenic organism is one that involves the transfer of genetic material between different species, whereas the term cisgenic is sometimes used in cases where gene transfers occur within a species. Currently, most GM crops are transgenic; however, this may change as gene editing techniques are increasingly used.

4. Herbicide-tolerant plants make up approximately 70 percent of all GM plants, the majority of which confer tolerance to the herbicide glyphosate. Glyphosate interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthetase, which is present in all plants and is required for the synthesis of aromatic amino acids phenylalanine, tyrosine, and tryptophan. When applied to resistant crops, glyphosate kills weeds without killing the crops. It has been a more efficient, economical, and less damaging tool in weed control than traditional methods of mechanical weeding and repeated tillage.

6. Some measures include providing high-dose vitamin A supplements and encouraging growth of fresh fruits and vegetables, but success has been limited due to the costs involved. In the 1990s, scientists created Golden Rice, a genetically engineered crop that synthesizes the vitamin A precursor, beta-carotene. However, the beta-carotene produced can provide only 15—20 percent of the recommended daily requirement of vitamin A. Later versions of Golden Rice 2 involved the introduction of similar genes from maize that led to a much higher production of beta-carotene—estimated to supply the full childhood daily requirement of vitamin A. At present, Golden Rice 2 is being tested in preparation for use in Bangladesh and the Philippines.

8. The biolistic method of gene introduction achieves DNA transfer by coating the transforming DNA in a heavy metal to form particles that are fired at high speed into plant cells using a gene gun. The introduced DNA may migrate into the cell nucleus and integrate into a plant chromosome. Transformation using *Agrobacterium tumifaciens* results in a higher rate of transformation, since the genetically engineered Ti plasmid can integrate the cloned DNA directly into the plant genome at random sites.

10. Roundup-Ready soybeans are resistant to the herbicide glyphosate, a broad-spectrum herbicide that interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthetase (needed for the plant to synthesize the aromatic amino acids phenylalanine, tyrosine, and tryptophan). Two copies of the *epsps* gene were cloned from *Agrobacterium* strain CP4 and introduced into soybeans using biolistic bombardment.

Discussion Question Answer

2. There are many positions taken and bills filed in various states to address the question of GM food labeling. Generally, many feel a "right to know" would allow consumers to make educated choices about the food they consume. Consumers would consider it an advantage to be able to judge the safety of a given food if they had information about the possibility that it contains GM components. Others wonder about the usefulness of a GM label if there is little information provided as to how the food has been modified. Of what value would it be to know that food was genetically modified if the science and specifics about the modifications were not included? How much background knowledge would be needed by the consumer to be able to interpret such information?

SPECIAL TOPIC 5

Review Question Answers

2. In *ex vivo* gene therapy, a potential genetic correction takes place in cells that have been removed from the patient. *In vivo* gene therapy treats cells of the body through the introduction of DNA into the patient.

4. In many cases, therapeutic DNA hitches a ride with genetically engineered viruses, such as retrovirus or adenovirus vectors. Nonviral delivery methods may use chemical assistance to cross cell membranes, nanoparticles, or cell fusion with artificial vesicles.

6. White blood cells, T cells in this case, were used because they are key players in the mounting of an immune response, which Ashanti was incapable of developing. A normal copy of the *ADA* gene was engineered into a retroviral vector, which then infected many of her T cells. Those cells that expressed the *ADA* gene were then injected into Ashanti's bloodstream, and some of them populated her bone marrow. At the time of Ashanti's treatment, targeted gene therapy was not possible, so integration of the *ADA* gene into Ashanti's genome probably did not replace her defective gene.

8. Gene editing involves the removal, correction, or replacement of a defective gene, whereas traditional gene therapy involves the addition of a therapeutic gene that coexists with the defective copy. Gene editing can alter one or several bases of a gene or replace the gene entirely. To some extent, gene editing is designed to alleviate one of the major pitfalls of gene therapy, random DNA integration.

10. ZFNs, or zinc-finger nucleases, consist of a DNA cutting domain from the restriction endonuclease, *Fok*I and a DNA binding domain containing a zinc-finger motif. The DNA-binding domain can be engineered to recognize any sequence. The nuclease portion of a ZFN provides a DNA cutting property that may eventually allow for targeted gene therapy.

Discussion Question Answers

2. Generally, gene therapy is an accepted procedure, given appropriate conditions, for the relief of genetic disease states. Since it is a fairly expensive medical approach, considerable debate attends its use. It remains to be seen whether insurance companies will embrace what might be considered experimental treatments. Use of gene therapy to enhance the competitive status of individuals (genetic enhancement or gene doping) is presently viewed as cheating by most organizations and the public. It is unlikely that germ-line therapy will be viewed favorably by the public or scientific communities; however, this and other issues mentioned here will be the subject of considerable future debate.

4. Germ-line and embryo therapy is controversial for a number of reasons. In addition to those given in the text are concerns about human experimentation—especially experimentation involving embryos. Ethical concerns aside, there are also safety concerns to consider. Among these concerns are the following: The treatment would need to be shown to be effective and efficient. Side effects, specifically effects of mistargeting, would need to be documented, quantified, and addressed. Further, altered viability of treated germ cells and embryos would also need to be addressed.

SPECIAL TOPIC 6

Review Question Answers

2. Nancy Wexler and her team collected pedigree information as well as blood samples from some individuals in each pedigree. DNA isolated from these samples was used to determine genotypes for the donors (allowing James Gusella to establish linkage between the HD locus and the C haplotype) and was ultimately screened by exon trapping to locate the *HTT* gene.

4. These experiments showed that protein aggregation in the brain and the resulting motor symptoms could be reversed when the mutant *HTT* gene was silenced shortly after the appearance of motor symptoms. This suggested that the disease might be controlled or reversed if treated during early stages.

6. The presence of the first exon (which contains the CAG repeat region) has been shown to be sufficient to cause the abnormalities seen in HD. This suggests that the presence of an expanded repeat region is necessary and, possibly, sufficient.

8. Three molecular approaches have been tested in mice for potential use in humans. (1) Repression of *mHTT* transcription by gene editing using ZFNs: The ZFN constructs were delivered by injection into the striatal region of brain. (2) Gene silencing (degradation of target mRNA) using ASOs complementary to the mutant allele: ASOs were delivered by transfusion into the cerebrospinal fluid. This method is in Phase I clinical trials in humans. (3) Gene editing using CRISPR-Cas9. In trials in human cells, Cas9 and sgRNA that recognized PAM sites adjacent to the mutant allele, but missing from the normal allele, were used. In mice, constructs for Cas9 and for sgRNAs were injected into the striatum. In addition, drug therapies that allow refolding of mHTT into a less toxic form are also being explored.

Discussion Question Answers

2. Understanding the functions of the mutant HTT protein is critical in the development and assessment of treatment strategies. However, understanding the function of the normal HTT protein will also be important to avoid potential treatments that are not specific to the mutant protein but also inhibit the normal protein.

4. TNR diseases show genetic anticipation. In HD, anticipation is seen with paternal inheritance of the mutant allele and is more likely when the father inherited the allele maternally. It is believed this form of anticipation may be the link to epigenetic methylation that occurs during imprinting.

This page intentionally left blank

Glossary

7-methylguanosine (m^7G) cap A guanosine residue that has been modified with a methyl group (CH_3) at the 7 position of the base, and has been added to the 5′ end of a eukaryotic mRNA through a 5′ to 5′ triphosphate bridge. The cap is important for mRNA stability, mRNA export from the nucleus, and translation initiation.

A-DNA An alternative form of right-handed, double-helical DNA. Its helix is more tightly coiled than the more common B-DNA, with 11 base pairs per full turn. In the A form, the bases in the helix are displaced laterally and tilted in relation to the longitudinal axis.

accession number An identifying number or code assigned to a nucleotide or amino acid sequence for entry and cataloging in a database.

acrocentric chromosome A chromosome or chromosome fragment with no centromere.

activation domain (AD) The region of a transcription factor that interacts with transcriptional machinery to initiate transcription.

activators A class of transcription factors that bind to enhancers or proximal-promoter elements to increase the transcription of a target gene.

active site The substrate-binding site of an enzyme; in other proteins, the portion whose structural integrity is required for function.

additive variance (V_A) Genetic variance attributed to the substitution of one allele for another at a given locus. This variance can be used to predict the rate of response to phenotypic selection in quantitative traits.

alignment Sequences of DNA, RNA or amino acids are compared and aligned based on sequence similarities.

allele-specific oligonucleotides (ASO) Short nucleotide chains, synthesized in the laboratory, usually 15–20 bp in length, that under carefully controlled conditions will hybridize only to a perfectly matching complementary sequence.

allele One of the possible alternative forms of a gene, often distinguished from other alleles by phenotypic effects.

allopolyploidy Polyploid condition formed by the union of two or more distinct chromosome sets with a subsequent doubling of chromosome number.

allosteric effect A conformational change in the active site of a protein brought about by interaction with an effector molecule.

allotetraploid An allopolyploid containing two genomes derived from different species.

alternative splicing Generation of different protein molecules from the same pre-mRNA by incorporation of a different set and order of exons into the mRNA product.

aminoacyl (A) site One of the three sites, or pockets, in the large and small subunits of the ribosome that may be occupied by tRNA during translation. This is the first to be occupied by each charged tRNA corresponding to the triplet codon of an mRNA.

aminoacyl tRNA synthetases Enzymes that catalyze the attachment of an amino acid to the appropriate tRNA.

amniocentesis A procedure in which fluid and fetal cells are withdrawn from the amnion, a membrane surrounding the fetus; used for genetic testing of the fetus.

amphidiploid An autopolyploid condition composed of four copies of the same genome.

amplicon Region of DNA (or RNA) that results from duplication during DNA replication; also applies to DNA (or RNA) replicated by the polymerase chain reaction (PCR).

anabolism The metabolic synthesis of complex molecules from less complex precursors.

anaphase The stage of cell division (mitosis and meiosis) in which chromosomes begin moving to opposite poles of the cell.

aneuploidy A condition in which the chromosome number is not an exact multiple of the haploid set.

anneal Joining together of DNA fragments by hydrogen-bonding.

annotation Analysis of genomic nucleotide sequence data to identify the protein-coding genes, the non-protein-coding genes, and the regulatory sequences and function(s) of each gene.

anticodon In a tRNA molecule, the nucleotide triplet that binds to its complementary codon triplet in an mRNA molecule.

antigen A molecule, often a cell-surface protein, that is capable of eliciting the formation of antibodies.

antisense oligonucleotide (ASO) A short, single-stranded DNA or RNA molecule complementary to a specific sequence.

apoptosis A genetically controlled program of cell death, activated as part of normal development or as a result of cell damage.

Argonaute A family of proteins found within the RNA-induced silencing complex (RISC) with endonuclease activity associated with the destruction of target mRNAs.

artificial selection Selection practiced artificially during research efforts. See *selection*.

assisted reproductive technologies (ART) The set of technologies used to treat infertility and assist couples in achieving pregnancy.

attenuation A regulatory process in some bacterial operons that terminates transcription prematurely, thus reducing the production of the mRNA encoding the structural genes in that operon.

autism spectrum disorder (ASD) A range of symptoms and neurodevelopmental disorders characterized by specific social, cognitive, and communication behaviors or deficits.

autonomously replicating sequences (ARSs) Origins of DNA replication, about 100 nucleotides in length, found in yeast chromosomes. ARS elements are also present in organelle DNA.

autopolyploidy Polyploid condition resulting from the duplication of one diploid set of chromosomes.

autoradiography Production of a photographic image by radioactive decay. Used to localize radioactively labeled compounds within cells and tissues or to identify radioactive probes in various blotting techniques. See also *Southern blotting*.

auxotroph A mutant microorganism or cell line that through mutation has lost the ability to synthesize one or more substances required for growth.

B-DNA The conformation of DNA most often found in cells and which serves as the basis of the Watson–Crick double-helical model. There are 10 base pairs per full turn of its right-handed helix, with the nucleotides stacked 0.34 nm apart. The helix has a diameter of 2.0 nm.

bacterial artificial chromosomes (BACs) Cloning vectors derived from bacterial chromosomes; designed to replicate larger fragments of cloned DNA than plasmids.

bacteriophage A virus that infects bacteria, using it as the host for reproduction. Often referred to as a phage.

balancer chromosome Chromosome containing one or more inversions that suppress crossing over with its homolog and which carries a dominant marker that is usually lethal when homozygous.

Barr body Densely staining DNA-positive mass seen in the somatic nuclei of mammalian females. Discovered by Murray Barr, this body represents an inactivated X chromosome. Also called a *sex chromatin body*.

base analogs Purine or pyrimidine bases that differ structurally from one normally used in biological systems but whose chemical behavior is the same. For example, 5-bromouracil, which “looks like” thymidine, substitutes for it, and after incorporation into a DNA molecule, can lead to mutations.

base substitution A single base change in a DNA molecule that produces a mutation. There are two types of substitutions: *transitions*, in which a purine is substituted for a purine, or a pyrimidine for a pyrimidine; and *transversions*, in which a purine is substituted for a pyrimidine or vice versa.

basic leucine zipper (bZIP) motif A common protein domain in eukaryotic transcription factors that function as dimers. Leucine residues in one region mediate dimerization between two bZIP-containing proteins while regions with basic amino acids, such as arginine and lysine, mediate sequence-specific DNA binding.

binary switch genes These genes program cells to follow one of several alternative developmental pathways.

bioinformatics A field that focuses on the design and use of software and computational methods for the storage, analysis, and management of biological information such as nucleotide or amino acid sequences.

biotechnology Commercial and/or industrial processes that utilize biological organisms or products.

bipotential gonads In mammalian embryos, a precursor to gonads are the the gonadal or genital ridges, described as bipotential gonads because they can develop into male or female gonads based on genetic and hormonal influences.

bivalents Synapsed homologous chromosomes in the first prophase of meiosis.

BLAST (Basic Local Alignment Search Tool) Any of a family of search engines designed to compare or query nucleotide or amino acid sequences against sequences in databases. BLAST also calculates the statistical significance of the matches.

“blue-white” screening DNA cloning technique used to distinguish host bacterial cells containing recombinant plasmids (white colonies) from host cells containing nonrecombinant plasmids (blue colonies).

Bombay phenotype A rare variant of the ABO antigen system in which affected individuals do not have A or B antigens and thus appear to have blood type O, even though their genotype may carry unexpressed alleles for the A and/or B antigens.

broad-sense heritability (H^2) The contribution of the genotypic variance responsible for the phenotypic variation of a trait observed in a population.

bromodeoxyuridine (BrdU) A mutagenically active base analog of thymidine in which the methyl group at the 5′ position in thymine is replaced by bromine.

β-**galactosidase** A bacterial enzyme, encoded by the *lacZ* gene, which converts lactose into galactose and glucose.

β-**globin** Polypeptide subunit of hemoglobin. A combination of two alpha and two beta chains comprise hemoglobin.

β-**thalassemia** Inherited blood disorder caused by a reduction or absence of β-hemoglobin subunits.

CAAT box A highly conserved DNA sequence found in the promoter region of eukaryotic genes upstream of the transcription start site. This sequence is recognized and bound by transcription factors.

cancer stem cells Tumor-forming cells in a cancer that can give rise to all the cell types in a particular form of cancer. These cells have the properties of normal stem cells: self-renewal and the ability to differentiate into multiple cell types.

capillary electrophoresis A group of analytical methods that separates large and small charged molecules in a capillary tube by their size-to-charge ratio. Detection and analysis of separated components takes place in the capillary usually by use of ultraviolet (UV) light.

carcinogens Physical or chemical agents that cause cancer.

Cas9 A CRISPR-associated nuclease from a type II CRISPR-Cas system that directs target DNA cleavage. Cas9 from the bacterium *Streptococcus pyogenes* is a well-studied protein that has been adapted as a tool for genome editing in eukaryotes.

catabolism A metabolic reaction in which complex molecules are broken down into simpler forms, often accompanied by the release of energy.

catabolite repression The selective inactivation of an operon by a metabolic product of the enzymes encoded by the operon.

catabolite-activating protein (CAP) A catabolite-activating protein; a protein that binds cAMP and regulates the activation of inducible operons.

***cdc* mutation** A class of *c*ell *d*ivision *c*ycle (*cdc*) mutations in yeast that affects the timing of and progression through the cell cycle.

cell cycle The sequence of growth phases in a cell; divided into G0, G1 (gap I), S (DNA synthesis), G2 (gap II), and M (mitosis). A cell may temporarily or permanently withdraw from the cell cycle, in which case it is said to enter the G0 stage.

cell theory The theory that all organisms are made of cells and that all cells come from pre-existing cells.

cell-cycle checkpoints Regulated transitions from one stage to another during the cell cycle.

CEN region The DNA region of centromeres critical to their function. In yeasts, fragments of chromosomal DNA, about 120 bp in length, that when inserted into plasmids confer the ability to segregate during mitosis.

central dogma of molecular genetics The classical concept that genetic information flow progresses from DNA to RNA to proteins. Although exceptions are now known, this idea is central to an understanding of gene function.

centriole A cytoplasmic organelle composed of nine groups of microtubules, generally arranged in triplets. Centrioles function in the generation of cilia and flagella and serve as foci for the spindles in cell division.

centromere The specialized heterochromatic chromosomal region at which sister chromatids remain attached after replication, and the site to which spindle fibers attach to the chromosome during cell division. The location of the centromere determines the shape of the chromosome during the anaphase portion of cell division. Also known as the primary constriction.

centrosome The region of the cytoplasm containing a pair of centrioles.

chance deviation The inherent error in a predictive statistical model, occurring strictly due to chance.

chaperone A protein that facilitates the folding of a polypeptide into the three-dimensional shape of a functional protein.

checkpoints See *cell-cycle checkpoints*.

chi-square (χ^2) analysis A statistical test to determine whether or not an observed set of data is equivalent to a theoretical expectation.

chiasmata (sing., chiasma) The crossed strands of nonsister chromatids seen in the first meiotic division. Regarded as the cytological evidence for exchange of chromosomal material, or crossing over.

chloroplast A self-replicating cytoplasmic organelle containing chlorophyll. The site of photosynthesis.

chorionic villus sampling (CVS) A technique of prenatal diagnosis in which chorionic fetal cells are retrieved and used to detect cytogenetic and biochemical defects in the embryo.

chromatin immunoprecipitation (ChIP) An analytical method used to identify DNA-binding proteins that bind to DNA sequences of interest. In ChIP, antibodies to specific proteins are used to isolate DNA sequences that bind these proteins.

chromatin The complex of DNA, RNA, histones, and nonhistone proteins that make up uncoiled chromosomes, characteristic of the eukaryotic interphase nucleus.

chromatin remodeling complexes Large multi-subunit enzyme complexes that use the energy of ATP hydrolysis to move and rearrange nucleosomes within chromatin.

chromatin remodeling A process in which the structure of chromatin is altered by a protein complex, resulting in changes in the transcriptional state of genes in the altered region.

chromomere A coiled, bead-like region of a chromosome, most easily visualized during cell division. The aligned chromomeres of polytene chromosomes are responsible for their distinctive banding pattern.

chromosomal mutation An alteration of the genome of an organism resulting in the duplication, deletion, or rearrangements of the diploid chromosomal content of an organism. Also called a chromosomal aberration.

chromosomal theory of inheritance The idea put forward independently by Walter Sutton and Theodor Boveri that chromosomes are the carriers of genes and the basis for the Mendelian mechanisms of segregation and independent assortment.

chromosome In bacteria, a DNA molecule containing the organism's genome; in eukaryotes, a DNA molecule complexed with RNA and proteins to form a threadlike structure containing genetic information arranged in a linear sequence; a structure that is visible during mitosis and meiosis.

chromosome-banding technique Technique for the differential staining of mitotic or meiotic chromosomes to produce a characteristic banding pattern; or selective staining of certain chromosomal regions such as centromeres, the nucleolus organizer regions, and GC- or AT-rich regions.

chromosome conformation capture (3C) A technique used to determine the spatial organization of chromatin with a eukaryotic cell's nucleus.

chromosome territory The discrete region that a chromosome occupies in the eukaryotic nucleus.

circular RNAs (circRNAs) RNAs with a circular single-stranded conformation. Some of these act as "molecular sponges" that bind to and block the action of microRNAs that modulate gene expression.

***cis*-acting DNA element**

***cis*-acting DNA element** A DNA sequence that regulates the expression of a gene located on the same chromosome. This contrasts with a *trans*-acting element where regulation is under the control of a sequence on the homologous chromosome.

cisgenic A genetically modified organism that contains a genetic material that was transferred from a member of the same species.

cistron That portion of a DNA molecule (a gene) that encodes a single-polypeptide chain; defined by a genetic test as a region within which two mutations cannot complement each other.

classical genetics See *forward genetics*

clone Identical molecules, cells, or organisms derived from a single ancestor by asexual or parasexual methods; for example, a DNA segment that has been inserted into a plasmid or chromosome of a phage or a bacterium and replicated to produce many copies, or an organism with a genetic composition identical to that used in its production.

cloning vectors See *vector*.

coactivators Proteins that do not directly bind to DNA but that work together with transcriptional activators to activate the transcription of a target gene.

coding strand In a double-stranded DNA molecule, the strand opposite and complementary to the template strand, which is not transcribed during transcription. So named because it is very similar in sequence to the RNA transcript.

codominance A condition in which the phenotypic effects of a gene's alleles are fully and simultaneously expressed in the heterozygote.

codon A triplet of messenger RNA (mRNA) nucleotides that specifies a particular amino acid or a start or stop signal in the genetic code. Sixty-one codons specify the amino acids used in proteins, and three codons, called stop codons (UAG, UAA, UGA), signal termination of growth of the polypeptide chain. One codon (AUG) acts as a start codon in addition to specifying an amino acid.

coefficient of coincidence (*C*) A ratio of the observed number of double crossovers divided by the expected number of such crossovers.

coefficient of inbreeding (*F*) The probability that two alleles present in a zygote are descended from a common ancestor.

cohesin A protein complex that holds sister chromatids together during mitosis and meiosis and facilitates attachments of spindle fibers to kinetochores.

colchicine An alkaloid compound that inhibits spindle formation during cell division. In the preparation of karyotypes, it is used for collecting a large population of cells inhibited at the metaphase stage of mitosis.

colinearity The linear relationship between the nucleotide sequence in a gene (or the RNA transcribed from it) and the order of amino acids in the polypeptide chain specified by the gene.

Combined DNA Index System (CODIS) A standardized set of 13 short tandem repeat (STR) DNA sequences used by law enforcement and government agencies in preparing DNA profiles.

comparative genomics Comparing genomes from different organisms to evaluate genetic and evolutionary similarity and other elements.

competence In bacteria, the transient state or condition during which the cell can bind and internalize exogenous DNA molecules, making transformation possible.

competing endogenous RNAs (ceRNAs) RNAs expressed in the cell, usually lncRNAs, that serve as decoys for the binding of miRNAs thus enabling the miRNA target genes to be expressed.

compiling Assembly of a complete, final reference sequence by combining data from multiple sequencing runs.

complementarity Chemical affinity between nitrogenous bases of nucleic acid strands as a result of hydrogen bonding. Responsible for the base pairing between the strands of the DNA double helix and between DNA and RNA strands during gene expression in cells and during the use of molecular hybridization techniques.

complementary DNA (cDNA) libraries A collection of cloned cDNA sequences.

complementation group An array of mutations that all test negatively when assayed in a complementation test. Thus, they are alleles in the same gene.

complex trait A quantitative trait whose phenotypic variation is the result of the interaction of additive alleles from multiple genes and environmental factors. Also called a *multifactorial trait*.

computer-automated high-throughput DNA sequencing Computer automated DNA sequencing technique that can produce large amounts (high-throughput) of DNA sequence in relatively short periods of time compared to manual sequencing.

concordance Condition when identical twins both express a trait or neither of them express that trait.

conditional mutation A mutation expressed only under a certain condition; that is, a wild-type phenotype is expressed under certain (permissive) conditions and a mutant phenotype under other (restrictive) conditions.

conjugation Temporary fusion of two single-celled organisms for the sexual transfer of genetic material.

consanguineous Related by a common ancestor within the previous few generations.

consensus sequence The sequence of nucleotides in DNA or amino acids in proteins most often present in a particular gene or protein.

constitutive mutations Mutations in bacterial operons that results in the continuous transcription of the structural genes in the operon.

contigs A continuous DNA sequence reconstructed from overlapping DNA sequences derived by cloning or sequence analysis.

contiguous fragments See *contigs*.

continuous variation A phenotype variation in which quantitative traits range from one phenotypic extreme to another in an overlapping or continuous fashion.

copy number variation (CNV) Any DNA segments larger than 1 kb that are repeated a variable number of times in the genome.

core enzyme The subunits of an enzyme necessary for catalytic activity. For DNA polymerase III, the core enzyme consists of three subunits: alpha, beta, and theta that confer catalytic activity to the holoenzyme.

core promoter The minimum part of a promoter needed for accurate initiation of transcription. Contains several core-promoter elements spanning a region of approximately 80 nucleotides including the transcription start site.

correlation coefficient (*r*) In quantitative genetic studies, a statistical value describing the degree of association between two interrelated traits.

covariance In quantitative genetic studies, a statistical value describing how much observed variation is common to two interrelated traits

CpG island A short region of regulatory DNA found upstream of genes that contain unmethylated stretches of sequence with a high frequency of C and G nucleotides.

CRISPR (clustered regularly interspaced palindromic repeats)-Cas system The adaptive immunity mechanism present in many bacteria, which utilizes CRISPR RNAs to guide Cas nucleases to invading complementary DNAs and destroy them. The CRISPR-Cas mechanism has also been exploited to introduce specific mutations in eukaryotic genomes.

CRISPR-associated (*cas*) genes Genes located physically near the CRISPR locus in bacteria and archaea that are necessary for at least one of the three steps of the CRISPR-Cas mechanism: spacer acquisition, crRNA biogenesis, and target interference.

CRISPR-derived RNAs (crRNAs) RNA transcripts from a CRISPR locus in a bacteria or archaea genome are processed into a mature crRNA, which then guides a Cas nuclease (such as Cas9) to cleave complementary DNA sequences.

crossing over The exchange of chromosomal material (parts of chromosomal arms) between homologous chromosomes by breakage and reunion. The exchange of material between nonsister chromatids during meiosis is the basis of genetic recombination.

crRNA biogenesis One of the steps in the CRISPR-Cas mechanism in which RNAs are transcribed and processed by Cas proteins.

cyclic adenosine monophosphate (cAMP) An important regulatory molecule in both bacteria and eukaryotes.

cyclins In eukaryotic cells, a class of proteins that are synthesized and degraded in synchrony with the cell cycle and regulate passage through stages of the cycle.

cytokinesis The division or separation of the cytoplasm at the end of cell division (mitosis and meiosis).

cytoplasmic polyadenylation element (CPE) A *cis*-regulatory sequence in some eukaryotic mRNA 3′ UTRs that regulates the length of the poly-A tail in the cytoplasm, and thus the translation of the mRNA.

cytoplasmic polyadenylation element binding protein (CPEB) An RNA-binding protein that binds to the cytoplasmic polyadenylation element in some mRNA 3′ UTRs and regulates poly-A tail length and translation of the mRNA.

dCas9 A Cas9 protein with engineered amino acid substitutions that render the nuclease domains “dead” or inactive such it can still bind a DNA sequence complementary to a sgRNA, but will not cut the DNA.

deadenylases Eukaryotic enzymes that degrade the poly-A tail of an mRNA in a 3′ to 5′ direction.

deadenylation-dependent decay A eukaryotic mRNA degradation pathway that is initiated by shortening the poly-A tail of an mRNA and subsequent degradation by either the exosome in the 3′ to 5′ direction or decapping of the mRNA and degradation in the 5′ to 3′ direction by the XRN1 exoribonuclease.

deadenylation-independent decay A eukaryotic mRNA degradation pathway that is initiated by decapping of an mRNA followed by degradation of the mRNA in the 5′ to 3′ direction by the XRN1 exoribonuclease.

decapping enzymes Eukaryotic enzymes that remove the m^7G cap at the 5′ end of an mRNA.

deletion A chromosomal mutation, also referred to as a deficiency, involving the loss of chromosomal material.

determination Establishment of a specific pattern of gene activity and developmental fate for a given cell, usually prior to any manifestation of the cell's future phenotype.

development The attainment of a differentiated state by the cells of an organism, except for stem cells.

dideoxy chain-termination sequencing One of the first DNA sequencing techniques; also referred to as Sanger sequencing. Technology relies on modified nucleotides called dideoxynucleotides (ddNTPs), which terminate a newly synthesized strand of DNA when incorporated during a sequencing reactions.

dideoxynucleotide A nucleotide containing a deoxyribose sugar lacking a hydroxyl group. It stops further chain elongation when incorporated into a growing polynucleotide and is used in the Sanger method of DNA sequencing.

differentiation The complex process of change by which cells and tissues attain their adult structure and functional capacity.

dihybrid cross A genetic cross involving two characters in which the parents possess different forms of each character (e.g., yellow, round × green, wrinkled peas).

diploid The condition when cells contain homologous pairs of each chromosome, one derived from the paternal parent and one from the maternal parent.

directional selection A selective force that changes the frequency of an allele in a given direction, either toward fixation or toward elimination.

discontinuous variation A pattern of variation for a trait whose phenotypes fall into two or more distinct classes.

discordance Condition when one member of a set of identical twins expresses a trait, while the other member does not.

disjunction The separation of chromosomes during the anaphase stage of cell division.

disruptive selection Simultaneous selection for phenotypic extremes in a population, usually resulting in the production of two phenotypically discontinuous strains.

dizygotic twins Twins produced from separate fertilization events; two ova fertilized independently. Also known as *fraternal twins.*

DNA fingerprinting A molecular method for identifying an individual member of a population or species. A unique pattern of DNA fragments is obtained by restriction enzyme digestion followed by Southern blot hybridization using minisatellite probes. See also *DNA profiling; STR sequences*.

DNA gyrase One of a class of enzymes known as topoisomerases that converts closed circular DNA to a negatively supercoiled form prior to replication, transcription, or recombination. The enzyme acts during DNA replication to reduce molecular tension caused by supercoiling.

DNA helicase An enzyme that participates in DNA replication by unwinding the double helix near the replication fork.

DNA knockout technology Generation of a *null mutation* in a gene that is subsequently introduced into an organism using transgenic techniques, causing a loss of function in the targeted gene. Often used in mice. See also *gene targeting*.

DNA libraries Collections of cloned DNA in host cells; cDNA or genomic libraries are common types of libraries. Can be screened to identify particular genes of interest.

DNA ligase An enzyme that forms a covalent bond between the 5′ end of one polynucleotide chain and the end of another polynucleotide chain. It is also called polynucleotide-joining enzyme.

DNA methylation the enzymatically controlled process of transferring a methyl group from a donor molecule to a base in DNA.

DNA microarray analysis An ordered arrangement of DNA sequences or oligonucleotides on a substrate (often glass). Microarrays are used in quantitative assays of DNA—DNA or DNA—RNA binding to measure profiles of gene expression (for example, during development or to compare the differences in gene expression between normal and cancer cells).

DNA polymerase An enzyme that catalyzes the synthesis of DNA from deoxyribonucleotides utilizing a template DNA molecule.

DNA profiling A method for identification of individuals that uses variations in the length of short tandem repeating (STR) DNA sequences that are widely distributed in the genome.

DNA transposons Mobile genetic elements that are major components of many eukaryotic and prokaryotic genomes. They are excised from one site in the genome and inserted into another, often causing mutations.

DNA typing In law enforcement, the use of a standard set of short tandem repeats (STRs) to identify an individual. Also known as *forensic DNA fingerprinting*. See also *DNA profiling*.

DNA-binding domain (DBD) A region of a transcription factor that recognizes and binds to a specific DNA sequence, such as an enhancer, silencer, or proximal-promoter element, to influence the transcription of a target gene.

dominant mutation A mutation in one allele that confers a mutant phenotype in a diploid organism, even in the presence of a wild-type allele

dominant-negative mutation A mutation whose gene product acts in opposition to the normal gene product, usually by binding to it to form dimers.

dosage compensation A genetic mechanism that equalizes the levels of expression of genes at loci on the X chromosome. In mammals, this is accomplished by random inactivation of one X chromosome, leading to Barr body formation.

driver mutation A mutation in a cancer cell that contributes to tumor progression.

Drosha A nuclear RNA-specific nuclease involved in the maturation of microRNAs. It removes 5′ and 3′ non-self-complementary regions of a primary miRNA to produce a pre-miRNA.

duplication A chromosomal aberration in which a segment of the chromosome is repeated.

electrophoresis A technique that separates a mixture of molecules by their differential migration through a stationary medium (such as a gel) under the influence of an electrical field.

ELSI program A program established by the National Human Genome Research Institute in 1990 as part of the Human Genome Project to sponsor research on the ethical, legal, and social implications of genomic research and its impact on individuals and social institutions.

embryonic stem (ES) cells Cells derived from the inner cell mass of early blastocyst mammalian embryos. These cells are pluripotent, meaning they can differentiate into any of the embryonic or adult cell types characteristic of the organism.

Encyclopedia of DNA Elements (ENCODE) Project A worldwide consortium of researchers using experimental approaches and bioinformatics to identify and analyze functional elements of the genome (such as transcriptional start sites, promoters, and enhancers) that regulate expression of human genes.

endonuclease An enzyme that hydrolyzes internal phosphodiester bonds in a single- or double-stranded polynucleotide chain.

endoplasmic reticulum (ER) A membranous organelle system in the cytoplasm of eukaryotic cells. In rough ER, the outer surface of the membranes is ribosome-studded; in smooth ER, it is not.

endopolyploidy The increase in chromosome sets within somatic nuclei that results from endomitotic replication.

endoribonucleases Eukaryotic enzymes that cleave an RNA at an internal site creating new 5′ and 3′ ends.

endosymbiotic theory The proposal that self-replicating cellular organelles such as mitochondria and chloroplasts were originally free-living bacterial organisms that entered into a symbiotic relationship with eukaryotic cells.

enhancement gene therapy Gene therapy for the purpose of enhancing a desired trait and not for disease treatment or prevention.

enhanceosome A complex of transcriptional activators and coactivators that directs the transcriptional activation of a target gene.

enhancer A DNA sequence that enhances transcription and the expression of structural genes. Enhancers can act over a distance of thousands of base pairs and can be located upstream, downstream, or internal to the gene they affect, differentiating them from promoters.

environmental genomics See *metagenomics*

enzyme A protein or complex of proteins that catalyzes a specific biochemical reaction by lowering the energy of activation that would otherwise be required to initiate the reaction.

epigenesis The idea that an organism or organ arises through the sequential appearance and development of new structures, in contrast to *preformationism,* which holds that development is the result of the assembly of structures already present in the egg.

epigenetics The study of the effects of reversible chemical modifications to DNA and/or histones on the pattern of gene expression. Epigenetic modifications do not alter the nucleotide sequence of DNA.

epigenome The set of chemical modifications made to DNA and histones that are present in each cell at a specific time period.

epimutations Heritable changes in gene expression that are associated with changes in the pattern of DNA methylation and not with any changes in the DNA sequence.

episome In bacterial cells, a circular genetic element that can replicate independently of the bacterial chromosome or integrate into and replicate as part of the chromosome.

epistasis The nonreciprocal interaction between nonallelic genes such that one gene influences or interferes with the expression of another gene, leading to a specific phenotype.

Ethical, Legal, Social Implications (ELSI) Program A program established by the National Human Genome Research Institute in 1990 as part of the Human Genome Project to sponsor research on the ethical, legal, and social implications of genomic research and its impact on individuals and social institutions.

euchromatin Chromatin or chromosomal regions that are lightly staining and relatively uncoiled during the interphase portion of the cell cycle. Euchromatic regions contain most of the structural genes.

eukaryotes Organisms having true nuclei and membranous organelles and whose cells divide by mitosis and meiosis.

euploidy A condition in which a cell has a chromosome number that is an exact multiple of the haploid number.

excision repair Removal of damaged DNA segments followed by repair. Excision can include the removal of individual bases (base excision repair) or of a stretch of damaged nucleotides (nucleotide excision repair).

exit (E) site One of the three sites, or pockets, in the large and small subunits of the ribosome that may be occupied by tRNA during translation. This is the site where an uncharged tRNA, which has already contributed its amino acid to the growing polypeptide chain, leaves the ribosome.

exome sequencing A DNA-sequencing method in which only the protein-coding regions (exons) of the genome are sequenced.

exon shuffling A molecular mechanism for the evolution of new genes whereby genetic recombination leads to the duplication or deletion of exons from a gene, or leads to exons from different genes being brought together to encode a new gene product.

exons The DNA segments of a gene that contain the sequences that, through transcription and translation, are eventually represented in the final polypeptide product.

exonuclease An enzyme that breaks down nucleic acid molecules by breaking the phosphodiester bonds at the 3′- or 5′-terminal nucleotides.

exoribonucleases Eukaryotic enzymes that degrade RNA via the removal of terminal nucleotides.

exosome complex A eukaryotic enzyme with exoribonuclease activity that degrades mRNAs in the 3′ to 5′ direction.

expression QTLs (eQTLs) Genomic loci that affect expression of one or more genes involved in a quantitative trait.

expression vectors Plasmids or phages carrying promoter regions designed to cause expression of inserted DNA sequences.

expressivity The degree to which a phenotype for a given trait is expressed.

extranuclear inheritance Transmission of traits by genetic information contained in the DNA of cytoplasmic organelles such as mitochondria and chloroplasts.

F^- cell A bacterial cell that does not contain a fertility factor and that acts as a recipient in bacterial conjugation.

F^+ cell A bacterial cell that contains a fertility factor and that acts as a donor in bacterial conjugation.

F_1 generation The first filial generation; the progeny resulting from the first cross in a series.

F_2 generation The second filial generation; the progeny resulting from a cross of the F_1 generation.

F factor An episomal plasmid in bacterial cells that confers the ability to act as a donor in conjugation.

F pilus On bacterial cells possessing an F factor, a filament-like projection that plays a role in conjugation.

FISH See *fluorescence* in situ *hybridization.*

fitness A measure of the relative survival and reproductive success of a given individual or genotype.

fluorescence *in situ* hybridization (FISH) A method of *in situ* hybridization that utilizes probes labeled with a fluorescent tag, causing the site of hybridization to fluoresce when viewed using ultraviolet light.

folded-fiber model A model of eukaryotic chromosome organization in which each sister chromatid consists of a single chromatin fiber composed of double-stranded DNA and proteins wound together like a tightly coiled skein of yarn.

forensic DNA fingerprinting In law enforcement, the use of a standard set of short tandem repeats (STRs) to identify an individual. Also known as *DNA typing*. See also *DNA profiling*.

forensic science The use of laboratory scientific methods to obtain data used in criminal and civil law cases.

forward genetics The classical approach used to identify a gene controlling a phenotypic trait in the absence of knowledge of the gene's location in the genome or its DNA sequence. Accomplished by isolating mutant alleles and mapping the gene's location, most traditionally using recombination analysis. Once mapped, the gene may be cloned and further studied at the molecular level. An approach contrasted with *reverse genetics*.

founder effect A form of genetic drift. The establishment of a population by a small number of individuals whose genotypes carry only a fraction of the different alleles present in the parental population.

fragile site A heritable gap, or nonstaining region, of a chromosome that can be induced to generate chromosome breaks.

frameshift mutation A mutational event leading to the insertion or deletion (indels) of a number of base pairs in a gene that is not a multiple of three. This shifts the codon reading frame in all codons that follow the mutational site.

fraternal twins See *dizygotic (DZ) twins*.

functional genomics Analysis of DNA sequence data to propose functions for sequenced DNA such as protein-coding and non-coding genes, regulatory element etcs.

G0 stage A nondividing but metabolically active state (G-zero) that cells may enter from the G1 phase of the cell cycle.

G1 checkpoint A point in the G1 phase of the cell cycle when a cell becomes committed to initiating DNA synthesis and continuing the cycle or withdrawing into the G0 resting stage.

G1 (gap I) stage The phase during the cell cycle between G0 and the S phase, during which the cell develops and grows.

G2 (gap II) stage The phase during the cell cycle following the S phase, during which the cell, having replicated its DNA, prepares for mitosis.

gain-of-function mutation A type of mutation in which the gene product takes on a new function and produces a phenotype different from that of the normal allele and from any loss-of-function alleles.

gamete A specialized reproductive cell with a haploid number of chromosomes.

gap genes Genes expressed in contiguous domains along the anterior—posterior axis of the *Drosophila* embryo that regulate the process of segmentation in each domain.

GC box In eukaryotes, a region in a promoter containing a 5′-GGGCGG-3′ sequence, which is a binding site for transcriptional regulatory proteins.

gene The fundamental physical unit of heredity, whose existence can be confirmed by allelic variants and which occupy a specific chromosomal locus. A DNA sequence coding for a single polypeptide or an RNA molecule.

gene amplification The process by which gene sequences are selected and differentially replicated either extrachromosomally or intrachromosomally.

gene chips See *DNA microarray analysis*.

gene editing Using engineered enzymes and specialized editing systems such as CRISPR-Cas to precisely modify the sequence of a particular gene.

gene family A number of closely related genes derived from a common ancestral gene by duplication and sequence divergence over evolutionary time.

gene flow The exchange of genes between two populations; brought about by the dispersal of gametes resulting from the migration of individuals between populations.

gene interaction The production of novel phenotypes by the interaction of alleles of different genes.

gene knockout Generation of a *null mutation* in a gene that is subsequently introduced into an organism using transgenic techniques, causing a loss of function in the targeted gene. Often used in mice. See also *gene targeting*.

gene pool The total of all alleles possessed by the reproductive members of a population.

gene redundancy The presence of several genes in an organism's genome that all have variations of the same function.

gene targeting A transgenic technique used to create and introduce a specifically altered gene into an organism. In mice, gene targeting often involves the induction of a specific mutation in a cloned gene that is subsequently introduced into the genome of a gamete involved in fertilization. The organism produced is bred to produce adults homozygous for the mutation, for example, the creation of a *gene knockout*.

gene therapy A therapeutic approach for providing a normal copy of a gene, replacement of a defective gene, or supplementing a gene for treating or curing a genetic disorder.

gene-regulatory networks (GRNs) The combination of binary switch genes and signaling pathways that direct the differentiation of specific tissue types and organs.

general transcription factors (GTFs) A class of DNA-binding proteins that bind to specific sites within a gene's promoter and are required to load RNA polymerase onto the DNA, and for the initiation of transcription.

genetic anticipation The phenomenon in which the severity of symptoms in genetic disorders increases from generation to generation and the age of onset decreases from generation to generation. It is caused by the expansion of trinucleotide repeats within or near a gene and was first observed in myotonic dystrophy.

genetic background The impact of the collective genome of an organism on the expression of a gene under investigation.

genetic code The deoxynucleotide triplets that encode the 20 amino acids or specify initiation or termination of translation.

genetic drift Random variation in allele frequency from generation to generation, most often observed in small populations.

genetic engineering The technique of altering the genetic constitution of cells or individuals by the selective removal, insertion, or modification of individual genes or gene sets.

genetically modified organism (GMO) A plant or animal whose genome has been altered in ways that do not occur naturally, most often using recombinant DNA and the techniques of genetic engineering.

genetics The branch of biology concerned with the study of inherited variation. More specifically, the study of the origin, transmission, and expression of genetic information.

genic balance theory In *Drosophila*, sex is determined by the ratio of X-chromosomes and autosomes.

genome The set of hereditary information encoded in the DNA of an organism, including both the protein-coding and non—protein-coding sequences.

Genome 10K project Genomics project to sequence 10,000 vertebrate genomes.

genome-wide association study (GWAS) Analysis of genetic variation across an entire genome, searching for linkage (associations) between variations in DNA sequences and a genome region encoding a specific phenotype.

genomic analysis General term describing a range of approaches for studying the genome (all DNA in an organisms' cells).

genomic imprinting The process by which the expression of an allele depends on whether it has been inherited from a male or a female parent. Also referred to as parental imprinting.

genomic library A collection of clones that contains all the DNA sequences of an organism's genome.

genomic variation Individual changes or variations in the human (reference) genome

genomics A subdiscipline of genetics created by the union of classical and molecular biology with the goal of sequencing and understanding genes, gene interaction, genetic elements, as well as the structure and evolution of genomes.

genotype The allelic or genetic constitution of an organism; often, the allelic composition of one or a limited number of genes under investigation.

germ-line therapy Gene therapy involving germ cells or mature gametes as targets for gene transfer.

Goldberg–Hogness box A short nucleotide sequence 20–30 bp upstream from the initiation site of eukaryotic genes to which RNA polymerase II binds. The consensus sequence is TATAAAA. Also known as a *TATA box*.

gonadal (genital) ridges In mammalian embryos, a precursor to gonads are the gonadal or genital ridges, described as bipotential gonads because they can develop into male or female gonads based on genetic and hormonal influences.

gratuitous inducer A molecule such as IPGT that is a chemical analogue of lactose, which in the lactose operon in bacteria induces the transcription of the structural genes, but itself is not metabolized by the gene products of the operon.

green fluorescent protein (GFP) A naturally occurring protein from the jellyfish *Aequorea victoria* that emits green light when excited by blue to ultraviolet light. GFP is commonly used as a tool in genetics where green fluorescence is used as a marker for gene expression, protein localization, or to label specific cell types.

H substance The carbohydrate group present on the surface of red blood cells to which the A and/or B antigen may be added. When unmodified, it results in blood type O.

haploid number (*n*) The number of homologous chromosome pairs characteristic of an organism or species.

haploinsufficiency In a diploid organism, a condition in which an individual possesses only one functional copy of a gene with the other inactivated by mutation. The amount of protein produced by the single copy is insufficient to produce a normal phenotype, leading to an abnormal phenotype. In humans, this condition is present in many autosomal dominant disorders.

haplotypes A set of alleles from closely linked loci carried by an individual that are inherited as a unit.

Hardy–Weinberg law The principle that genotype frequencies will remain in equilibrium in an infinitely large, randomly mating population in the absence of mutation, migration, and selection.

harlequin chromosomes Paired human sister chromatids stained to reveal sister chromatid exchanges.

helix–turn–helix (HTH) motif In DNA-binding proteins, the structure of a region in which a turn of four amino acids contains two helices at right angles to each other.

hemizygous Having a gene present in a single dose in an otherwise diploid cell. Usually applied to genes on the X chromosome in heterogametic males.

hemoglobin (Hb) An iron-containing, oxygen-carrying multimeric protein occurring chiefly in the red blood cells of vertebrates.

heritability For a given trait, a measure of the proportion of total phenotypic variation in a population that is due to genetic factors.

heterochromatin The heavily staining, late-replicating regions of chromosomes that are prematurely condensed in interphase.

heteroduplex DNA molecules Regions of double-stranded DNA that are hybrids of two strands from different sources. The two strands may contain one or more different nucleotide sequences, or mismatches. Heteroduplexes can result from crossing-over during homologous recombination in meiosis.

heteroduplex A double-stranded nucleic acid molecule in which each polynucleotide chain has a different origin. It may be produced as an intermediate in a recombinational event or by the *in vitro* reannealing of single-stranded complementary molecules.

heterogametic sex The sex that produces gametes containing unlike sex chromosomes. In mammals, the male is the heterogametic sex.

heterogeneous nuclear ribonucleoproteins (hnRNPs) A complex of RNA and proteins in the nucleus that regulate mRNA splicing and export of mRNA from the nucleus.

heterogeneous nuclear RNA (hnRNA) The collection of RNA transcripts in the nucleus, consisting of precursors to and processing intermediates for rRNA, mRNA, and tRNA. Also includes RNA transcripts that will not be transported to the cytoplasm, such as snRNA.

heterogeneous trait A mutant phenotype that may occur as the result of a mutation in any one of many genes required for normal expression of the trait during development, e.g., hereditary deafness.

heterokaryon A somatic cell containing nuclei from two different sources.

heteromorphic chromosomes Dissimilar chromosome pairs that characterize one sex or the other in a wide range of species, e.g., XY in mammals and ZW in chickens.

heteroplasmy Variation in the DNA within organelles such as mitochondria and chloroplasts within the same cell.

heterozygote An individual with different alleles at one or more loci. Such individuals will produce unlike gametes and therefore will not breed true.

Hfr Strains of bacteria exhibiting a high frequency of recombination. These strains have a chromosomally integrated F factor that is able to mobilize and transfer part of the chromosome to a recipient F^- cell.

high-frequency recombination See *Hfr*.

high-throughput sequencing A collection of DNA-sequencing methods that outperform the standard (Sanger) method of DNA sequencing by a factor of 100–1000 and reduce sequencing costs by more than 99 percent. Also called *next-generation sequencing*.

histone acetyltransferase (HAT) A class of enzymes that catalyze the addition of an acetyl group to histone proteins, which leads to an open chromatin configuration that is permissive for gene expression.

histone code Chemical modifications of amino acids in histone tails (the N-terminal ends of histone molecules, projecting from nucleosomes). These modifications influence DNA–histone interactions and promote or repress transcription.

histone deacetylase (HDAC) A class of enzymes that catalyzes the removal of acetyl groups from histone proteins, which leads to a closed chromatin configuration that is refractory to gene expression.

histones Positively charged proteins complexed with DNA in the nucleus. They are rich in the basic amino acids arginine and lysine and function in coiling DNA to form nucleosomes.

Holliday structure In DNA recombination, an intermediate stage seen in transmission electron microscope images as an X-shaped structure showing four single-stranded DNA regions.

holoenzyme For proteins with multiple subunits, the complex formed by the union of all subunits necessary for all functions of the enzyme.

homeobox A sequence of 180 nucleotides that encodes a sequence of 60 amino acids called a *homeodomain*, which is part of a DNA-binding protein that acts as a transcription factor.

homeodomain The 60-amino acid region of a protein that binds DNA and is encoded by a 180-nucleotide sequence known as a homeobox.

homeotic mutation A mutation that causes a tissue normally determined to form a specific organ or body part to alter its pathway of differentiation to form another structure.

homeotic selector genes A family of eukaryotic transcription factors that contain a homeodomain and which specify the developmental fate of cells within each body segment.

homogametic sex The sex that produces gametes with identical sex-chromosome content; in mammals, the female is homogametic.

homologous chromosomes Chromosomes that synapse or pair during meiosis and that are identical with respect to their genetic loci and centromere placement.

homologous genes Genes related through evolution.

homologous recombination A type of recombination involving the exchange of nucleotide sequences between two similar or identical DNA molecules. Homologous recombination creates new combinations of DNA sequences during meiosis and is used in cells to repair double-stranded breaks.

homology-directed repair (HDR) A genomic DNA repair mechanism that is induced by double-strand DNA breaks. In this type of repair, an intact molecule of DNA, such as a homologous chromosome in diploid eukaryotes, is used as a template for repair of a damaged homolog

homozygote An individual with identical alleles for a gene or genes of interest. These individuals will produce identical gametes (with respect to the gene or genes in question) and will therefore breed true.

homunculus In the incorrect theory of preformationism, the idea that the egg or sperm contained a fully formed human called the homunculus, and that development consisted only of enlarging the preformed individual.

horizontal gene transfer The nonreproductive transfer of genetic information from one organism to another, across species and higher taxa (even domains). This mode is contrasted with vertical gene transfer, which is the transfer of genetic information from parent to offspring. In some species of bacteria and archaea, up to 5 percent of the genome may have originally been acquired through horizontal gene transfer.

hot spots Genome regions where mutations are observed with a high frequency. These include a predisposition toward single-nucleotide substitutions or unequal crossing over.

Human Genome Project International effort to identify all human genes and to sequence an estimated 3 billion based pairs of the entire human genome; also included goals to sequence genomes for model organisms; to evaluate genetic variation in humans; and to address ethical, legal and social issues among other goals.

Human Microbiome Project (HMP) National Insitutes of Health project to sequence the genomes of microorganisms that reside inside and on humans.

hybridization A cytological technique for pinpointing the chromosomal location of DNA sequences complementary to a given nucleic acid or polynucleotide.

hydrogen bond A weak electrostatic attraction between a hydrogen atom covalently bonded to an oxygen or a nitrogen atom and an atom that contains an unshared electron pair.

identical twins See *monozygotic twins*.

immunoglobulin (Ig) The class of serum proteins having the properties of antibodies.

immunotherapy Therapies that stimulate, suppress, or modify the actions of components of the patient's immune system, in order to treat diseases such as cancer.

imprinting See *genomic imprinting*.

inbreeding Mating between closely related organisms.

incomplete dominance Expressing a heterozygous phenotype that is distinct from the phenotype of either homozygous parent. Also called *partial dominance*.

induced mutations Mutations that result from the actions of exogenous agents, either natural or artificial.

inducible enzyme system An enzyme system under the control of an inducer, a regulatory molecule that acts to block a repressor and allow transcription.

initiation codon The nucleotide triplet AUG that in an mRNA molecule codes for incorporation of the amino acid methionine as the first amino acid in a polypeptide chain.

***in situ* hybridization** A cytological technique for pinpointing the chromosomal location of DNA sequences complementary to a given nucleic acid or polynucleotide.

insulator A DNA sequence that serves as a boundary element. Insulators are located between an enhancer and the promoter of a non-target gene to prevent the enhancer from influencing the transcription of the non-target gene.

interchromatin compartments The regions between chromosome territories in the nucleus that are largely void of DNA.

interference (*I*) A measure of the degree to which one crossover affects the incidence of another crossover in an adjacent region of the same chromatid. Negative interference increases the chance of another crossover; positive interference reduces the probability of a second crossover event.

interphase In the cell cycle, the interval between divisions.

intervening sequence See *intron*

intrinsic termination In bacteria, one mechanism by which transcription is terminated. Following the transcription of a stretch of U residues, the weak A-U base-pairing between the DNA template and the RNA leads to dissociation and transcription termination.

intron Any segment of DNA that lies between coding regions in a gene. Introns are transcribed but are spliced out of the RNA product and are not represented in the polypeptide encoded by the gene. Short for intervening sequence.

inversion A chromosomal aberration in which a chromosomal segment has been reversed.

inversion loop The chromosomal configuration resulting from the synapsis of homologous chromosomes, one of which carries an inversion.

isoforms A set of related proteins that are encoded by different alternative splice forms of the same gene.

karyokinesis The process of nuclear division.

karyotype The chromosome complement of a cell or an individual. An arrangement of metaphase chromosomes in a sequence according to length and centromere position.

kinases A broad class of enzymes that phosphorylate a substrate molecule such as a protein, nucleic acid, carbohydrate, or lipid.

kinetochore A protein structure that assembles on the centromere during mitosis and meiosis. It is the site of microtubule attachment during cell division.

knock-in animals See *transgenic animals*.

Kozak sequence A short nucleotide sequence adjacent to the initiation codon that is recognized as the translational start site in eukaryotic mRNA.

lagging strand During DNA replication, the strand synthesized in a discontinuous fashion, in the direction opposite of the replication fork. See also *Okazaki fragment*.

lampbrush chromosomes Meiotic chromosomes characterized by extended lateral. Although most intensively studied in amphibians, these structures occur in meiotic cells of organisms ranging from insects to humans.

leader sequence That portion of an mRNA molecule from the 5′ end to the initiating codon, often containing regulatory or ribosome-binding sites.

leading strand During DNA replication, the strand synthesized continuously in the direction of the replication fork.

linkage The condition in which genes are present on the same chromosome, causing them to be inherited as a unit, provided that they are not separated by crossing over during meiosis.

linking number (*L*) The number of times that two strands of a closed, circular DNA duplex cross over each other.

locus (pl., loci) The place on a chromosome where a particular gene is located.

long interspersed elements (LINEs) Long, repetitive sequences found interspersed in the genomes of higher organisms.

long noncoding RNAs (lncRNAs) RNAs longer than 200 nucleotides that are noncoding transcripts and have various functions including epigenetic modification of DNA and regulation of gene-specific transcription.

long terminal repeat (LTR) Identical sequences found at both ends of a retroviral DNA.

loss of heterozygosity The loss of a wild-type allele at a heterozygous locus that also contains a mutant allele. Most commonly, loss of heterozygosity occurs through deletion of a chromosomal region or a recombination event that converts the wild-type allele to the mutant allele sequence.

loss-of-function mutation A mutation that produces alleles encoding proteins with reduced or no function.

Lyon hypothesis The proposal that there is random inactivation of the maternal or paternal X chromosome in somatic cells of mammalian females early in development. All daughter cells will have the same X chromosome inactivated as in the cell they descended from, producing a mosaic pattern of expression of X chromosome genes.

lysis The disintegration of a cell brought about by the rupture of its membrane.

lysogeny The process by which the DNA of an infecting phage becomes repressed and integrated into the chromosome of the bacterial cell it infects.

MADS-box proteins A family of eukaryotic transcription factors characterized by the presence of a highly conserved 56-amino acid sequence DNA binding domain. In plants, MADS-box proteins play important roles in flower development.

male-specific region of the Y chromosome (MSY) Region of the Y-chromosome that does not recombine with the X-chromosome; contains a majority of genes specific to the Y-chromosome.

malignant tumors Tumors that have acquired to ability to spread to distant sites in the body and form new tumors.

mass spectrometry (MS) Technique that can be applied to proteomics and involves analyzing protein sample in gaseous forms to measure mass to charge ratios ions in a protein sample as as way to identify the amino acid sequence in a sample.

maternal effect Phenotypic effects in offspring attributable to genetic information in the oocyte derived from the maternal genome.

maternal-effect genes See *maternal effect*.

mean ($\bar{X}$) The arithmetic average.

Mediator During transcription initiation, this multiprotein complex serves as a coactivator and interacts with general transcription factors and RNA polymerase II.

meiosis The process of cell division in gametogenesis or sporogenesis during which the diploid number of chromosomes is reduced to the haploid number.

melting profile (T_m) The temperature at which a population of double-stranded nucleic acid molecules is half-dissociated into single strands.

meristic trait A phenotype in which quantitative variation can be measured in whole numbers.

merozygote A partially diploid bacterial cell containing, in addition to its own chromosome, a chromosome fragment introduced into the cell by transformation, transduction, or conjugation.

messenger RNA (mRNA) An RNA molecule transcribed from DNA and translated into the amino acid sequence of a polypeptide.

metacentric chromosome A chromosome with a centrally located centromere and therefore chromosome arms of equal lengths.

metafemale In *Drosophila*, a poorly developed female of low viability with a ratio of X chromosomes to sets of autosomes that exceeds 1.0. Previously called a *superfemale*.

metagenomics The study of DNA recovered from organisms collected from the environment as opposed to those grown as laboratory cultures. Often used for estimating the diversity of organisms in an environmental sample.

metamale In *Drosophila*, a poorly developed male of low viability with a ratio of X chromosomes to sets of autosomes that is below 0.5. Previously called a *supermale*.

metaphase The stage of cell division (mitosis and meiosis) in which condensed chromosomes lie in a central plane between the two poles of the cell and during which the chromosomes become attached to the spindle fibers.

metastases Secondary tumors that develop from cancer cells that migrate from the primary tumor and spread to distant sites in the body.

metastasis The process by which cancer cells spread from the primary tumor and establish malignant tumors in other parts of the body.

methylation Enzymatic transfer of methyl groups from S-adenosylmethionine to biological molecules, including phospholipids, proteins, RNA, and DNA. Methylation of DNA is associated with the regulation of gene expression and with epigenetic phenomena such as imprinting.

methylome the pattern of nucleic acid methylation present at a particular time in a genome or a specific cell type.

microfilaments Actin-containing microfibers that are a part of the structural framework of the cytoplasm.

microRNA (miRNA) Single-stranded RNA molecules approximately 20–23 nucleotides in length that regulate gene expression by participating in the degradation of mRNA.

microsatellite A short, highly polymorphic DNA sequence of 1–4 base pairs, widely distributed in the genome, that are used as molecular markers in a variety of methods. Also called *simple sequence repeats (SSRs).*

microtubules Fibers composed of tubulin, bunches of which are part of the structural framework of the cytoplasm (the cytoskeleton) and also which compose spindle fibers that facilitate chromosome migration during mitosis and meiosis.

minimal medium A medium containing only the essential nutrients needed to support the growth and reproduction of wild-type strains of an organism. Usually comprised of inorganic components that include a carbon and nitrogen source.

minisatellite Series of short tandem repeat sequences (STRs) 10–100 nucleotides in length that occur frequently throughout the genome of eukaryotes. Because the number of repeats at each locus is variable, the loci are known as variable number tandem repeats (VNTRs). Used in the preparation of DNA fingerprints and DNA profiles. See also *variable number tandem repeats (VNTRs); STR sequences*.

miRNA response elements (MREs) Sequences in RNAs that are complementary or partially complementary to miRNAs and thus serve as binding sites for miRNAs.

mismatch repair (MMR) A form of excision repair of DNA in which the repair mechanism is able to distinguish between the strand with the error and the strand that is correct.

missense mutation A mutation that changes a codon to that of another amino acid and thus results in an amino acid substitution in the translated protein. Such changes can make the protein nonfunctional.

mitochondria Self-reproducing, DNA-containing, cytoplasmic organelles in eukaryotes involved in generating the high-energy compound ATP. They are the so-called powerhouse of the cell.

mitochondrial DNA (mtDNA) Double-stranded, self-replicating circular DNA found in mitochondria that encodes mitochondrial ribosomal RNAs, transfer RNAs, and proteins used in oxidative respiratory functions of the organelle.

mitochondrial replacement therapy (MMR) The exchange of defective mitochondria in an egg with normal mitochondria to correct a mitochondrial genetic disorder. This is accomplished by transferring the nucleus from an egg with defective mitochondria into a nondefective egg that has had its nucleus removed. The reconstructed egg is fertilized via *in vitro* fertilization and implanted for development.

mitosis A form of cell division producing two progeny cells identical genetically to the parental cell—that is, the production of two cells from one, each having the same chromosome complement as the parent cell.

model organisms In genetics, model organisms are non-human species with well characterized genetics that are studied to understand basic biological processes. The expectation is that the process in the model organism can be extrapolated to other species, including humans.

molecular clock In evolutionary studies, a method that counts the number of differences in DNA or protein sequences as a way of measuring the time elapsed since two species diverged from a common ancestor.

molecular motors Specialized proteins that facilitate the movement of cellular components.

monoallelic expression (MAE) Process in which transcription occurs only from one of two homologous alleles in a diploid cell.

monohybrid cross A genetic cross involving only one character (e.g., $AA \times aa$).

monosomy An aneuploid condition in which one member of a chromosome pair is missing; having a chromosome number of $2n-1$.

monozygotic twins Twins produced from a single fertilization event; the first division of the zygote produces two cells, each of which develops into an embryo. Also known as *identical twins*.

mosaics Refers to individuals with cell types that contain different karyotypes (for example, in Turner syndrome, cells with 45,X and cells with 46,XY) and thus often show mixed phenotypes for a particular genetic condition compared with non-mosaics (for example, 45,X individuals with Turner syndrome).

mRNA See *messenger RNA*

Müllerian inhibiting substance (MIS) Protein produced by the developing mammalian testis; causes atrophy of embryonic tissue (Müllerian duct) that forms female reproductive tissues.

multifactorial trait See *complex trait*.

multigene families A set of genes descended from a common ancestral gene usually by duplication and subsequent sequence divergence. The globin genes are an example of a multigene family.

multiple alleles In a population of organisms, the presence of three or more alleles of the same gene.

multiple factor inheritance See *polygenic inheritance*.

multiple factor hypothesis Proposal describing the inheritance of a phenotypic character controlled by many genes, each behaving in a Mendelian fashion and contributing to the phenotype in a cumulative, or quantitative way.

mutagens Any agent that causes an increase in the rate of mutation.

mutation The process that produces an alteration in DNA or chromosome structure; in genes, the source of new alleles.

mutation rate The frequency with which mutations take place at a given locus or in a population.

mutator phenotype The high levels of genomic instability that occur in cancer cells.

N-formylmethionine (fMet) A molecule derived from the amino acid methionine by attachment of a formyl group to its terminal amino group. This is the first monomer used in the synthesis of all bacterial polypeptides. Also known as *N-formyl methionine.*

narrow-sense heritability The proportion of phenotypic variance in a population due to additive genotypic variance.

natural selection Differential reproduction among members of a species owing to variable fitness conferred by genotypic differences.

neoantigens Novel, nonself antigens present on cancer cells but not found on the patient's normal cells.

neutral mutation A mutation with no perceived immediate adaptive significance or phenotypic effect.

next-generation sequencing (NGS) technologies See *high-throughput DNA sequencing*.

noncoding RNAs (ncRNAs) RNAs that do not encode polypeptides.

nondisjunction A cell division error in which homologous chromosomes (in meiosis) or the sister chromatids (in mitosis) fail to separate and migrate to opposite poles; responsible for defects such as monosomy and trisomy.

nonhomologous end-joining (NHEJ) A genomic DNA repair mechanism that is induced by double-strand DNA breaks. This type of repair is error prone because broken ends of DNA molecules are randomly ligated together, which may lead to insertions, deletions, translocations, or inversions.

noninvasive prenatal genetic diagnosis (NIPGD) A noninvasive method of fetal genotyping that uses a maternal blood sample to analyze thousands of fetal loci using fetal cells or fetal DNA fragments present in the maternal blood.

nonsense codons The nucleotide triplets (UGA, UAG, and UAA) in an mRNA molecule that signal the termination of translation.

nonsense mutation A mutation that changes a codon specifying an amino acid into a termination codon, leading to premature termination during translation of mRNA.

nonsense-mediated decay (NMD) An eukaryotic mRNA degradation pathway that is triggered by the presence of a stop codon too far upstream of the poly-A tail of the mRNA.

nonsister chromatids Non-identical chromatids visible during mitosis and meiosis where each chromatid represents one or the other of the two members of a homologous pair of chromosomes. See *sister chromatids*.

NOR A chromosomal region containing the genes for ribosomal RNA (rRNA); most often found in physical association with the *nucleolus*.

normal distribution A probability function that approximates the distribution of random variables. The normal curve, also known as a Gaussian or bell-shaped curve, is the graphic display of a normal distribution.

Northern blot analysis Electrophoresis, blotting and hybridization technique for detecting and quantifying RNA expression.

Northern blotting An analytic technique in which RNA molecules are separated by electrophoresis and transferred by capillary action to a nylon or nitrocellulose membrane. Specific RNA molecules can then be identified by hybridization to a labeled nucleic acid probe.

Notch signaling pathway A highly conserved signaling pathway in multicellular organisms that determines cell fate during development.

nucleoid The DNA-containing region within the cytoplasm in bacterial cells.

nucleolar organizer region (NOR) A chromosomal region containing the genes for ribosomal RNA (rRNA); most often found in physical association with the *nucleolus*.

nucleolus The nuclear site of ribosome biosynthesis and assembly; associated with or formed in association with the DNA comprising the *nucleolar organizer region*.

nucleoside In nucleic acid chemical nomenclature, a purine or pyrimidine base covalently linked to a ribose or deoxyribose sugar molecule.

nucleosome In eukaryotes, a complex consisting of four pairs of histone molecules wrapped by two turns of a DNA molecule. The major structure associated with the organization of chromatin in the nucleus.

nucleotide In nucleic acid chemical nomenclature, a nucleoside covalently linked to one or more phosphate groups. Nucleotides containing a single phosphate linked to the 5′ carbon of the ribose or deoxyribose are the building blocks of nucleic acids.

nucleus The membrane-bound cytoplasmic organelle of eukaryotic cells that contains the chromosomes and nucleolus.

null allele A mutant allele that produces no functional gene product. Usually inherited as a recessive trait.

null hypothesis (H_0) Used in statistical tests, the hypothesis that there is no real difference between the observed and expected datasets. Statistical methods such as chi-square analysis are used to test the probability associated with this hypothesis.

nutrigenomics The study of how food and components of food affect gene expression.

Okazaki fragment The short, discontinuous strands of DNA produced on the lagging strand during DNA synthesis.

oligonucleotide A linear sequence of about 10–20 nucleotides connected by 5′-3′ phosphodiester bonds.

oncogene A gene whose activity promotes uncontrolled proliferation in eukaryotic cells. Usually a mutant gene derived from a *proto-oncogene*.

open reading frame (ORF) A nucleotide sequence organized as triplets that encodes the amino acid sequence of a polypeptide, including an initiation codon and a termination codon.

operator region In bacterial DNA, a region that interacts with a specific repressor protein to regulate the expression of an adjacent gene or gene set.

operon A genetic unit consisting of one or more structural genes encoding polypeptides, and an adjacent operator gene that regulates the transcriptional activity of the structural gene or genes.

ordered genetic code The pattern of triplet code sequences whereby chemically similar amino acids often share one or two middle bases.

orthologs Genes with sequence similarity found in two or more related species that arose from a single gene in a common ancestor.

p arm Shorter of the two arms extending from the centromere of the chromosome.

pair-rule genes Genes expressed as stripes around the blastoderm embryo during development of the *Drosophila* embryo.

palindrome In genetics, a sequence of DNA base pairs that read the same on complementary strands. Because the strands run antiparallel to one another in DNA, the base sequences on the two strands read the same backwards and forward when read from the 5′ end. For example:

5′-GAATTC-3′
3′-CTTAAG-5′

Palindromic sequences are noteworthy as recognition and cleavage sites for restriction endonucleases.

pangenome An attempt to show or display a genome inclusive of known major genome variations (as opposed to a reference genome that displays the most commonly observed sequence without variations).

paracentric inversion A chromosomal inversion that does not include the region containing the centromere.

paralogs Two or more genes in the same species derived by duplication and subsequent divergence from a single ancestral gene.

parental gamete A gamete whose chromosomes have undergone no genetic recombination.

partial dominance Expressing a heterozygous phenotype that is distinct from the phenotype of either homozygous parent. Also called *incomplete dominance*.

passenger mutations Mutations that accumulate in cancer cells that have no direct contribution to the development or progression of the cancer.

pedigree In human genetics, a diagram showing the ancestral relationships and transmission of genetic traits over several generations in a family.

penetrance The frequency, expressed as a percentage, with which individuals of a given genotype manifest at least some degree of a specific mutant phenotype associated with a trait.

peptide bond The covalent bond between the amino group of one amino acid and the carboxyl group of another amino acid.

peptidyl (P) site The middle of three sites, or pockets, in the large and small subunits of the ribosome that may be occupied by tRNA during translation. The growing polypeptide chain is attached to tRNA in this site and during each step of translation, peptide bond formation occurs, extending the polypeptide by one amino acid.

pericentric inversion A chromosomal inversion that involves both arms of the chromosome and thus the centromere.

Personal Genome Project A project to enroll 100,000 individuals to share their genome sequence, personal information, and medical history with researchers and the general public to increase understanding of the contribution of genetic and environmental factors to genetic traits.

personal genomics Whole-genome sequencing of individual genomes.

personalized medicine Devising and applying unique and specific therapies for each individual, based on the individual's unique molecular profiles. A part of precision medicine.

pharmacogenomics The study of how genetic variation influences the action of pharmaceutical drugs in individuals.

phenotype The overt appearance of a genetically controlled trait.

Philadelphia chromosome The product of a reciprocal translocation in humans that contains the short arm of chromosome 9, carrying the *C-ABL* oncogene, and the long arm of chromosome 22, carrying the *BCR* gene.

phosphatases A broad class of enzymes that remove a phosphate group from a substrate molecule such as a protein, nucleic acid, carbohydrate, or lipid.

phosphodiester bond In nucleic acids, covalent bonds by which a phosphate group links adjacent nucleotides, extending from the 5′ carbon of one pentose sugar (ribose or deoxyribose) to the 3′ carbon of the pentose sugar in the neighboring nucleotide. Phosphodiester bonds create the backbone of nucleic acid molecules.

photolyase See *photoreactivation enzyme (PRE)*.

photoreactivation enzyme (PRE) An enzyme that cleaves the cross-linking bonds in thymine dimers. Also called *photolyase*.

plaque On an otherwise opaque bacterial lawn, a clear area caused by the growth and reproduction of a single bacteriophage.

plaque assay A quantitative assay using a serial dilution of a solution containing a large unknown number of bacteriophages that determines the original density (phage per ml).

plasmid An extrachromosomal, circular DNA molecule that replicates independently of the host chromosome.

pleiotropy A condition in which a single mutation causes multiple phenotypic effects.

pluripotent See *totipotent*.

point mutation A mutation that can be mapped to a single locus. At the molecular level, a mutation that results in the substitution of one nucleotide for another.

polar body Produced in females at the first and second meiotic divisions, a discarded cell that contains one of the nuclei resulting from the division process, but lacking in in most ooplasm as a result of unequal cytokinesis.

poly-A binding protein A protein that binds to the poly-A tail on the 3′ end of eukaryotic mRNAs to stabilize them, help them be exported from the nucleus, and initiate translation.

poly-A polymerase An enzyme that catalyzes the addition of adenosine residues to the 3′ end of a eukaryotic mRNA to synthesize the poly-A tail.

poly-A ribonuclease (PARN) A eukaryotic enzyme that is recruited to some mRNAs in the cytoplasm to shorten the poly-A tail and thus downregulate translation.

polyadenylation signal sequence A conserved AAUAAA sequence in eukaryotic mRNAs. Once transcribed, the nascent transcript is cleaved roughly 10-35 base pairs downstream.

polycistronic mRNA A messenger RNA molecule that encodes the amino acid sequence of two or more polypeptide chains in adjacent structural genes. Characteristic of bacteria.

Polygenic inheritance The transmission of a phenotypic trait whose expression is dependent on an additive effect of many genes. Also referred to as *quantitative inheritance*.

polymerase chain reaction (PCR) A method for amplifying DNA segments that depends on repeated cycles of denaturation, primer annealing, and DNA polymerase—directed DNA synthesis.

polynucleotide A linear sequence of 20 or more nucleotides, joined by 5′-3′ phosphodiester bonds. See also *oligonucleotide*.

polypeptide A molecule composed of amino acids linked together by covalent peptide bonds. This term is used to denote the amino acid chain before it assumes its functional three-dimensional configuration and is called a protein.

polyploidy A condition in which a cell or individual has more than two haploid sets of chromosomes.

polyribosome A structure composed of two or more ribosomes associated with an mRNA engaged in translation. Also called a *polysome*.

polytene chromosome Literally, a many-stranded chromosome; one that has undergone numerous rounds of DNA replication without separation of the replicated strands, which remain in exact parallel register. The result is a giant chromosome with aligned chromomeres displaying a characteristic banding pattern, often studied in *Drosophila* larval salivary gland cells.

population A local group of actually or potentially interbreeding individuals belonging to the same species.

position effect A change in expression of a gene associated with a change in the gene's location within the genome.

positional cloning The identification and subsequent cloning of a gene without knowledge of its polypeptide product or function. The process uses cosegregation of mutant phenotypes with DNA markers to identify the chromosome containing the gene; the position of the gene is identified by establishing linkage with additional markers.

posttranscriptional modification Changes made to pre-mRNA molecules during conversion to mature mRNA. These include the addition of a methylated cap at the 5′ end and a poly-A tail at the 3′ end, excision of introns, and exon splicing.

posttranslational modification The processing or modification of the translated polypeptide chain by enzymatic cleavage, addition of phosphate groups, carbohydrate chains, or lipids.

postzygotic isolation mechanism A factor that prevents or reduces inbreeding by acting after fertilization to produce nonviable, sterile hybrids or hybrids of lowered fitness.

pre-initiation complex (PIC) Prior to transcription initiation in eukaryotes, this complex of RNA polymerase II and general transcription factors loads onto the promoter.

pre-miRNAs A stage of miRNA processing that occurs after an miRNA gene is transcribed into a primary miRNA and then is cleaved by Drosha to remove 5′ and 3′ ends.

precision medicine An individualized approach to disease diagnosis and treatment, which uses various molecular profiles to select precise therapies and to develop new treatments.

preformationism The idea that development is the result of the assembly of structures already present in the egg, in contrast to *epigenesis*, which holds that an organism or organ arises through the sequential appearance and development of new structures.

preimplantation genetic diagnosis (PGD) The removal and genetic analysis of unfertilized oocytes, polar bodies, or single cells from an early embryo (3–5 days old).

prezygotic isolation mechanism A factor that reduces inbreeding by preventing courtship, mating, or fertilization.

Pribnow box In bacterial genes, a 6-bp sequence to which the sigma (σ) subunit of RNA polymerase binds, upstream from the beginning of transcription. The consensus sequence for this box is TATAAT. Also referred to as the -10 site.

primary miRNAs (pri-miRNAs) Transcriptional product of a microRNA gene containing a 5′ methylated cap, a 3′ polyadenylated tail, and a hairpin structure derived from self-complementary sequences.

primary protein structure The sequence of amino acids in a polypeptide chain.

primary sex ratio (PSR) Ratio of males to females at fertilization, often expressed in decimal form (e.g., 1.06).

primer In nucleic acids, a short length of RNA or single-stranded DNA required for initiating synthesis directed by polymerases.

primordial germ cells (PGCs) Precursor cells in the early embryo that migrate into gonads and develop to form gametes.

prion An infectious pathogenic agent devoid of nucleic acid and composed of a protein, PrP, with a molecular weight of 27,000–30,000 Da. Prions are known to cause scrapie, a degenerative neurological disease in sheep; bovine spongiform encephalopathy (BSE, or mad cow disease) in cattle; and similar diseases in humans, including kuru and Creutzfeldt–Jakob disease.

proband An individual who is the focus of a genetic study leading to the construction of a pedigree tracking the inheritance of a genetically determined trait of interest. Formerly known as a *propositus*.

probe A macromolecule such as DNA or RNA that has been labeled and can be detected by an assay such as autoradiography or fluorescence microscopy. Probes are used to identify target molecules, genes, or gene products.

processing bodies (P bodies) Regions in the cytoplasm of the eukaryotic cell that contain many enzymes involved in mRNA decay. While many mRNAs that enter P bodies are degraded, some may be stored there for translation at a later time.

processivity The ability of an enzyme to carry out consecutive reactions before dissociating from its substrate. In the case of DNA polymerase III, processivity is the number of nucleotides added to a template strand before the enzyme falls off the template.

product law In statistics, the law holding that the probability of two independent events occurring simultaneously is equal to the product of their independent probabilities.

prokaryotes Organisms lacking a nuclear membrane, true chromosomes, and membranous organelles, e.g., bacteria and blue-green algae.

prometaphase Stage of cell division (mitosis and meiosis) during which the spindle fibers are assembled and attach to the centromeres of chromosomes, which begin their migration to the equatorial plate.

promoters A region of a gene where RNA polymerase binds and initiates transcription, located in the 5′ direction from the coding sequence.

proofreading A molecular mechanism for scanning and correcting errors in replication, transcription, or translation.

prophage A bacteriophage genome integrated into a bacterial chromosome that is replicated along with the bacterial chromosome. Bacterial cells carrying prophages are said to be *lysogenic* and to be capable of entering the *lytic cycle*, whereby phage particles are produced.

prophase The initial stage of cell division (mitosis and meiosis) during which the nuclear envelope breaks down, the nucleolus disintegrates, centrioles are formed and migrate to opposite ends of the cell, cytoplasmic microtubules are organized into spindle fibers, and diffuse chromatin fibers begin to condense into chromosomes.

protein A molecule composed of one or more polypeptides, each composed of amino acids covalently linked together. Proteins demonstrate *primary*, *secondary*, *tertiary*, and often, *quaternary structure*.

protein domain Amino acid sequences with specific conformations and functions that are structurally and functionally distinct from other regions of the same protein.

proteome The entire set of proteins expressed by a cell, tissue, or organism at a given time.

proteomics The study of the expressed proteins present in a cell at a given time.

proto-oncogene A gene that functions to initiate, facilitate, or maintain cell growth and division. Proto-oncogenes can be converted to *oncogenes* by mutation.

protoplast A bacterial or plant cell with the cell wall removed. Sometimes called a *spheroplast*.

protospacer adjacent motif (PAM) A short (2-6 basepair) DNA immediately adjacent to the DNA sequence that is targeted by a Cas nuclease of the CRISPR-Cas system. In the well-described CRISPR-Cas system of Streptococcus pyogenes, the PAM sequence is 5′-NGG-3′ immediately downstream of the target sequence on the non-complementary strand of the DNA.

prototroph A strain (usually of a microorganism) that is capable of growth on a defined, minimal medium. Wild-type strains are usually regarded as prototrophs and contrasted with *auxotrophs*.

pseudoautosomal region (PARs) A region on the human Y chromosome that is also represented on the X chromosome. Genes found in this region of the Y chromosome have a pattern of inheritance that is indistinguishable from genes on autosomes.

pseudogene A nonfunctional gene with sequence homology to a known structural gene present elsewhere in the genome. It differs from the functional version by insertions or deletions and by the presence of flanking direct-repeat sequences of 10–20 nucleotides.

puff A localized uncoiling and swelling in a polytene chromosome, usually regarded as a sign of active transcription.

pyrosequencing A high-throughput method of DNA sequencing that determines the sequence of a single-stranded DNA molecule by synthesis of a complementary strand. During synthesis, the sequence is determined by the chemiluminescent detection of pyrophosphate release that accompanies nucleotide incorporation into a newly synthesized strand of DNA.

q arm Longer of the two arms extending from the centromere of the chromosome.

QTL mapping population F_2 populations from crosses of inbred strains that carry different parts of the P_1 parental genomes and associated QTL genotypes and phenotypes.

quantitative inheritance See *polygenic inheritance*.

quantitative real-time PCR (qPCR) A variation of PCR (polymerase chain reaction) that uses fluorescent probes to quantitate the amount of DNA or RNA product present after each round of amplification.

quantitative trait loci (QTLs) Two or more genes that act on a single polygenic trait in a quantitative way.

quaternary protein structure Types and modes of interaction between two or more polypeptide chains within a protein molecule.

quorum sensing (QS) A mechanism used to regulate gene expression in response to changes in cellular population density.

R plasmid A bacterial plasmid that carries antibiotic resistance genes. Most R plasmids have two components: an r-determinant that carries the antibiotic resistance genes and the resistance transfer factor (RTF).

reactive oxygen species (ROS) Ions or molecules formed by incomplete reduction of oxygen. They include hydroxyl radicals, superoxides, and peroxides. ROS cause oxidative damage to DNA and other macromolecules.

reading frame A linear sequence of codons in a nucleic acid.

recessive mutation A mutation that results in a wil-type phenotype when present in a diploid organism and the other allele is also wild-type.

reciprocal cross A pair of crosses in which the genotype of the female in one is present as the genotype of the male in the other, and vice versa.

recognition sequence See *restriction site*.

recombinant DNA DNA molecules created by joining together pieces of DNA from different sources.

recombinant DNA technology A collection of methods used to create DNA molecules by *in vitro* ligation of DNA from two different organisms, and the replication and recovery of such recombinant DNA molecules.

recombinant gamete A gamete containing a new combination of alleles produced by crossing over during meiosis.

recombination The process that leads to the formation of new allele combinations on chromosomes.

reference genome Most commonly observed genome sequence based on whole genome sequencing of different individuals; serves as a baseline for genome analysis and comparison.

replication fork The Y-shaped region of a chromosome at the site of DNA replication.

replication The process whereby DNA is duplicated.

replicon The unit of DNA replication, beginning with DNA sequences necessary for the initiation of DNA replication. In bacteria, the entire chromosome is a replicon.

reporter gene A recombinant DNA tool that detects gene expression. The regulatory sequence of a gene of interest is fused to a coding sequence that confers an easily observable phenotype, such as fluorescence, and is inserted into an organism to learn when, where, and under what conditions the gene of interest is expressed.

rescue experiment An experimental approach that involves transferring a gene or protein to cells of an organism to restore a particular phenotype or function. Often used in experiments when a gene has been mutated. Investigators add back the relevant gene to try and rescue the phenotype and thus provide functional evidence for the gene being studied.

resistance transfer factor (RTF) A component of R plasmids that confers the ability to transfer the R plasmid between bacterial cells by conjugation.

restriction enzymes DNA cutting enzymes that cleave or “digest” DNA at specific sequences.

restriction enzymes See *restriction endonuclease*

restriction fragment length polymorphism (RFLP) Variation in the length of DNA fragments generated by restriction endonucleases. These variations are caused by mutations that create or abolish cutting sites for restriction enzymes. RFLPs are inherited in a codominant fashion and are extremely useful as genetic markers.

restriction map Physical map displaying the number, order and distances between restriction enzyme digestion sites in a particular segment of DNA.

restriction site A DNA sequence, often palindromic, recognized by a restriction endonuclease. The enzyme binds to and cleaves DNA at the restriction site.

retrotransposons Mobile genetic elements that are major components of many eukaryotic genomes. They are copied by means of an RNA intermediate and inserted at other chromosomal sites, often causing mutations.

retroviral vectors Gene delivery vectors derived from retroviruses. Often used in gene therapy.

retrovirus A type of virus that uses RNA as its genetic material and employs the enzyme reverse transcriptase during its life cycle.

reverse genetics An experimental approach used to discover gene function after the gene has been identified, isolated, cloned, and sequenced. The cloned gene may be knocked out (e.g., by *gene targeting*) or have its expression altered (e.g., by *RNA interference* or *transgenic overexpression*) and the resulting phenotype studied. An approach contrasted with *forward genetics*.

reverse transcriptase A polymerase that facilitates reverse transcription using RNA as a template to transcribe a single-stranded DNA molecule.

rho-dependent termination In bacteria, one mechanism by which transcription is terminated. The rho protein breaks DNA-RNA base pairing with its helicase activity, which leads to RNA dissociation and transcription termination.

ribonucleoprotein (RNP) particles or granules A collection of small, stable RNA molecules complexed with one or more proteins. RNPs are found in the nucleus, cytoplasm, and mitochondria; they have a variety of functions, including RNA splicing, transport, posttranscriptional regulation of mRNA, and protein export.

ribosomal RNA (rRNA) The RNA molecules that are the structural components of the ribosomal subunits. In bacteria, these are the 16*S*, 23*S*, and 5*S* molecules; in eukaryotes, they are the 18*S*, 28*S*, and 5*S* molecules. See also *Svedberg coefficient unit (S)*.

ribosome A ribonucleoprotein organelle consisting of two subunits, each containing RNA and protein molecules. Ribosomes are the site of translation of mRNA codons into the amino acid sequence of a polypeptide chain.

riboswitch An RNA-based intracellular sensor that binds to a small ligand, such as a metabolite, modulating control of gene expression.

ribozymes RNAs that catalyze specific biochemical reactions.

RNA editing Alteration of the nucleotide sequence of an mRNA molecule after transcription and before translation. There are two main types of editing: substitution editing, which changes individual nucleotides, and insertion/deletion editing, in which individual nucleotides are added or deleted.

RNA interference (RNAi) Inhibition of gene expression in which a protein complex (RNA-induced silencing complex, or RISC) containing a complementary (or partially complementary) RNA strand binds to an mRNA, leading to degradation or reduced translation of the mRNA.

RNA polymerase An enzyme that catalyzes the formation of an RNA polynucleotide strand using the base sequence of a DNA molecule as a template.

RNA sequencing Technique for determining the nucleotide sequence of RNA molecules.

RNA splicing The processing of a nascent transcript of RNA by the removal of introns and the joining together of exons.

RNA-binding proteins (RBPs) A class of proteins that bind to specific RNA sequences or RNA secondary structures and influence many posttranscriptional regulatory mechanisms such as alternative splicing, RNA decay, RNA stability, RNA transport and localization, and translation.

RNA-induced silencing complex (RISC) A protein complex containing an Argonaute family protein with endonuclease activity. siRNAs and miRNAs guide RISC to complementary mRNAs to cleave them.

Robertsonian translocation A chromosomal aberration created by breaks in the short arms of two acrocentric chromosomes followed by fusion of the long arms of these chromosomes at the centromere. Also called *centric fusion*.

S phase The "synthesis" portion of the cell cycle following the G1 phase during which DNA is replicated.

Sanger sequencing DNA sequencing by synthesis of DNA chains that are randomly terminated by incorporation of a nucleotide analog (dideoxynucleotides) followed by sequence determination by analysis of resulting fragment lengths in each reaction.

satellite DNA DNA that forms a minor band when genomic DNA is centrifuged in a cesium salt gradient. This DNA usually consists of short sequences repeated many times in the genome.

secondary protein structure The α-helical or β-pleated-sheet formations in a polypeptide, dependent on hydrogen bonding between certain amino acids.

secondary sex ratio The ratio of males to females at birth, usually expressed in decimal form (e.g., 1.05).

segment polarity genes Genes that regulate the spatial pattern of differentiation within each segment of the developing *Drosophila* embryo.

segmentation genes One or more classes of genes expressed in the early embryo that divide the embryo into a series of regions or segments.

segregation The separation of maternal and paternal homologs of each homologous chromosome pair into gametes during meiosis.

selectable marker gene Marker gene such as one that encodes a gene for antibiotic resistance allowing specific cells to be chosen or selected by phenotype or characteristics (for example, antibiotic resistance) provided by the marker gene

selection The changes that occur in the frequency of alleles and genotypes in populations as a result of differential reproduction.

selfing In plant genetics, the fertilization of a plant's ovules by pollen produced by the same plant. Reproduction by self-fertilization.

semiconservative replication A mode of DNA replication in which a double-stranded molecule replicates in such a way that the daughter molecules are each composed of one parental (old) and one newly synthesized strand.

severe combined immunodeficiency A collection of genetic disorders characterized by a lack of immune system response; both cell-mediated and antibody-mediated responses are missing.

sex chromatin body See *Barr body*.

sex ratio See *primary sex ratio* and *secondary sex ratio*.

sex-determining region Y (*SRY*) Essential gene on the human Y chromosome that controls male sexual development; encodes TDF protein.

sex-influenced inheritance A phenotypic expression conditioned by the sex of the individual. A heterozygote may express one phenotype in one sex and an alternate phenotype in the other sex (e.g., pattern baldness in humans).

sex-limited inheritance A trait that is expressed in only one sex even though the trait may not be X-linked or Y-linked.

Shine–Dalgarno sequence The nucleotides AGGAGG that serve as a ribosome-binding site in the leader sequence of bacterial genes. The 16*S* RNA of the small ribosomal subunit contains a complementary sequence to which the mRNA binds.

short interspersed elements (SINEs) Repetitive sequences found in the genomes of higher organisms. The 300-bp *Alu* sequence is a SINE element.

short tandem repeats (STRs) Short tandem repeats 2–9 base pairs long found within minisatellites. These sequences are used to prepare DNA profiles in forensics, paternity identification, and other applications.

short tandem repeats See *STR sequences*.

shotgun cloning The cloning of random fragments of genomic DNA into a vector (a plasmid or phage), usually to produce a library from which clones of specific interest can be selected for use, as in sequencing.

shugoshins A class of proteins involved in maintaining cohesion of the centromeres of sister chromatids during mitosis and meiosis.

sigma (σ) factor In RNA polymerase, a polypeptide subunit that recognizes the DNA binding site for the initiation of transcription.

signal transduction An intercellular or intracellular molecular pathway by which an external signal is converted into a functional biological response.

silencers A DNA sequence that reduces or blocks the transcription and the expression of genes. Silencers can act over a distance of thousands of base pairs and can be located upstream, downstream, or internal to the gene they affect.

silent mutation A mutation that alters the sequence of a codon but does not result in a change in the amino acid at that position in the protein.

single guide RNA (sgRNA) An engineered hybrid RNA molecule that combines sequences of the crRNA and tracrRNA of type II CRISPR-Cas systems into a single RNA. This makes CRISPR-Cas-mediated genome editing more convenient because two separate components are integrated into one.

single-nucleotide polymorphism (SNP) A variation in one nucleotide pair in DNA, as detected during genomic analysis. Present in at least 1 percent of a population, a SNP is useful as a genetic marker.

single-stranded binding proteins (SSBs) In DNA replication, proteins that bind to and stabilize the single-stranded regions of DNA that result from the action of unwinding proteins.

sister chromatid exchange (SCE) A crossing-over event in meiotic or mitotic cells involving the reciprocal exchange of chromosomal material between sister chromatids joined by a common centromere. Such exchanges can be detected cytologically after BrdU incorporation into the replicating chromosomes.

sister chromatids A pair of identical chromatids visible during mitosis and meiosis that are formed following replication of DNA of one member of a homologous chromosome pair.

sliding clamp loader A component of the DNA polymerase III holoenzyme consisting of five subunits that attach a circular protein complex to the polymerase in an ATP-dependent reaction.

sliding DNA clamp A component of the DNA polymerase III holoenzyme composed of multiple copies of the beta subunit that forms a circular structure attached to the polymerase, which promotes processivity.

small interfering RNAs (siRNAs) Small (or short) interfering RNAs. Short 20–25 nucleotide double-stranded RNA sequences with two 3′ overhanging nucleotides; they are processed by Dicer and participate in transcriptional and/or posttranscriptional mechanisms of gene regulation.

small noncoding RNAs (sRNAs) Any of a number of short RNAs that are noncoding transcripts that associate with the RNA-induced silencing complex (RISC) to regulate transcription or to regulate mRNAs posttranscriptionally.

small nuclear ribonucleoproteins (snRNPs) Abundant species of small RNA molecules ranging in size from 90 to 400 nucleotides that in association with proteins form RNP particles known as snRNPs or *snurps*. Located in the nucleoplasm, snRNAs have been implicated in the processing of pre-mRNA and may have a range of cleavage and ligation functions.

small nuclear RNA (snRNA) Abundant species of small RNA molecules ranging in size from 90 to 400 nucleotides that in association with proteins form RNP particles known as snRNPs or *snurps*. Located in the nucleoplasm, snRNAs have been implicated in the processing of pre-mRNA and may have a range of cleavage and ligation functions.

somatic cell hybridization A technique involving the fusion of somatic cells first utilized in the 1960s to assign human genes to their respective chromosomes.

somatic gene therapy Gene therapy involving somatic cells as targets for gene transfer.

somatic mutation A nonheritable mutation occurring in a somatic cell.

Southern blot analysis Electrophoresis, blotting and hybridization technique for detecting specific DNA fragments; pioneered by Ed Southern.

Southern blotting Developed by Edwin Southern, a technique in which DNA fragments produced by restriction enzyme digestion are separated by electrophoresis and transferred by capillary action to a nylon or nitrocellulose membrane. Specific DNA fragments can be identified by hybridization to a complementary radioactively labeled nucleic acid probe using the technique of *autoradiography*.

spacer acquisition One of the steps in the CRISPR-Cas mechanism in which spacer DNA sequences, often derived from invading bacteriophage genomes, are inserted into the CRISPR locus of a host bacterial or archaea genome.

spacer DNA DNA sequences found between genes. Usually, these are repetitive DNA segments.

speciation The process by which new species of plants and animals arise.

species A group of actually or potentially interbreeding individuals that is reproductively isolated from other such groups.

specification The mechanisms by which the developmental fate of embryonic cells are determined.

spectral karyotype A display of all the chromosomes in an organism as a karyotype with each chromosome stained in a different color.

spheroplast See *protoplast*.

spindle fibers Cytoplasmic fibrils formed during cell division that attach to and are involved with separation of chromatids at the anaphase stage of mitosis and meiosis as well as their movement toward opposite poles in the cell.

splice acceptor An AG dinucleotide sequence that defines the 3′ end of an intron during RNA splicing.

splice donor A GU dinucleotide sequence that defines the 5′ end of an intron during RNA splicing.

spliceforms Different mRNAs produced from the same gene through the process of alternative splicing.

spliceopathies A class of human diseases caused by defects in RNA splicing.

spliceosome The nuclear macromolecule complex within which splicing reactions occur to remove introns from pre-mRNAs.

splicing enhancers *Cis*-acting sequences in an mRNA that influence alternative splicing by promoting splicing at a nearby splice site.

splicing silencers *Cis*-acting sequences in an mRNA that influence alternative splicing by inhibiting splicing at a nearby splice site.

spontaneous generation The incorrect idea that living organisms can be created directly from nonliving material rather than by descent from other living organisms.

spontaneous mutation A random mutation that is not induced by a mutagenic agent.

spore A unicellular body or cell encased in a protective coat. Produced by some bacteria, plants, and invertebrates, spores are capable of surviving in unfavorable environmental conditions and give rise to a new individual upon germination. In plants, spores are the haploid products of meiosis.

SR proteins A class of Serine (S) and Arginine (R) repeat-containing proteins that influence mRNA splicing.

Src A protein kinase that phosphorylates many target proteins to regulate their activity, for example, regulating the activity of the RNA-binding protein ZBP1.

stabilizing selection Preferential reproduction of individuals with genotypes close to the mean for the population. A selective elimination of genotypes at both extremes.

standard deviation (*s*) A quantitative measure of the amount of variation present in a sample of measurements from a population calculated as the square root of the variance.

standard error of the mean Stastistical calculation that compares variation between the means of multiple samples in a population.

stone-age genomics Sequencing and analysis of ancient DNA samples

STR sequences Short tandem repeats 2-9 bases long found within minisatellite DNA sequences. Used in preparation of profiles in DNA forensics and other applications.

structural gene A gene that encodes the amino acid sequence of a polypeptide chain.

structural genomics A gene that encodes the amino acid sequence of a polypeptide chain.

submetacentric chromosome A chromosome with the centromere placed so that one arm of the chromosome is slightly longer than the other.

sum law The law that holds that the probability of one of two mutually exclusive outcomes occurring, where that outcome can be achieved by two or more events, is equal to the sum of their individual probabilities.

supercoiled DNA See *supercoiling*.

supercoiling In reference to the tertiary structure of DNA, the underwinding (creating negative supercoils) or overwinding (creating positive supercoils) that occurs when the helix is strained.

suppressor mutation A second mutation that reverts or relieves the effects of a previous mutation. Suppressor mutations can occur either within the same gene that contained the first mutation, or elsewhere in the genome.

Svedberg coefficient *(S)* A unit of measure for the rate at which particles (molecules) sediment in a centrifugal field. This rate is a function of several physicochemical properties, including size and shape. A rate of 1×10^{-13} seconds is defined as 1 Svedberg coefficient unit.

synapsis The pairing of homologous chromosomes at meiosis.

synkaryon The fusion of two gametic or somatic nuclei. Also, in somatic cell genetics, the product of nuclear fusion.

synteny testing Utilized during somatic cell hybridization analysis, the correlation of the presence or absence of each human chromosome with the presence or absence of a specific gene product.

synthetic biology A scientific discipline that combines science and engineering to research the complexity of living systems and to construct biological-based systems that do not exist in nature.

synthetic genome A genome assembled from chemically synthesized DNA fragments that is transferred to a host cell without a genome.

systems biology A scientific discipline that identifies and analyzes gene and protein networks to gain an understanding of intracellular regulation of metabolism, intra- and intercellular communication, and complex interactions within, between, and among cells.

TALENs See *transcription activator-like effector nucleases*.

target interference One of the steps in the CRISPR-Cas mechanism in which crRNAs guide a Cas nuclease to cleave target DNA sequences.

TATA box See *Goldberg—Hogness box*

tautomeric shift A reversible isomerization in a molecule, brought about by a shift in the location of a hydrogen atom. In nucleic acids, tautomeric shifts in the bases of nucleotides can cause changes in other bases at replication and are a source of mutations.

telocentric chromosome A chromosome in which the centromere is located at its very end.

telomerase The enzyme that adds short, tandemly repeated DNA sequences to the ends of eukaryotic chromosomes.

telomeres The heterochromatic terminal regions of a chromosome.

telophase The stage of cell division (mitosis and meiosis) in which the daughter chromosomes have reached the opposite poles of the cell and reverse the stages characteristic of prophase, re-forming the nuclear envelopes and uncoiling the chromosomes. Telophase ends during cytokinesis, which divides the cytoplasm and splits the parental cell into two daughter cells.

temperate phage A bacteriophage that can become a prophage, integrating its DNA into the chromosome of the host bacterial cell and making the latter lysogenic.

temperature-sensitive mutation A conditional mutation that produces a mutant phenotype at one temperature and a wild-type phenotype at another.

template strand In a double stranded DNA molecule, the strand that is transcribed by RNA polymerase during transcription.

termination factor, rho (ρ) In bacterial rho-dependent transcriptional termination, the rho protein breaks DNA-RNA base pairing with its helicase activity, which leads to RNA dissociation and transcription termination.

tertiary protein structure The three-dimensional conformation of a polypeptide chain in space, specified by the polypeptide's primary structure. The tertiary structure achieves a state of maximum thermodynamic stability.

testcross A cross between an individual whose genotype at one or more loci may be unknown and an individual who is homozygous recessive for the gene or genes in question.

testis-determining factor (TDF) Protein encodes by *SRY* gene; causes bipotential gonads of an embryo to form testes.

tetrad The four chromatids that make up paired homologs in the prophase of the first meiotic division. In eukaryotes with a predominant haploid stage (some algae and fungi), a tetrad denotes the four haploid cells produced by a single meiotic division.

third-generation sequencing (TGS) DNA-sequencing technologies based largely on methods for sequencing individual molecules of single-stranded DNA.

threshold effect Condition where normal phenotypic expression occurs any time a minimal level of wild type gene product is attained.

threshold traits Quantitative traits distinguished by displaying a number of discrete phenotypic classes.

Ti plasmid A bacterial plasmid used as a vector to transfer foreign DNA to plant cells.

topoisomerase A class of enzymes that converts DNA from one topological form to another, leading to *topoisomers*. During replication, a topoisomerase, *DNA gyrase*, facilitates DNA replication by reducing molecular tension caused by supercoiling upstream from the *replication fork*.

totipotent The capacity of a cell or an embryo part to differentiate into all cell types characteristic of an adult. This capacity is usually progressively restricted during development. Used interchangeably with *pluripotent*.

***trans*-acting factor** A gene product (usually a diffusible protein or an RNA molecule) that acts to regulate the expression of a target gene.

transactivating crRNA (tracrRNA) A small noncoding RNA that is a necessary component of many bacterial type II CRISPR-Cas systems during the crRNA biogenesis and target interference steps.

transcription activator-like effector nucleases (TALENs) Artificial DNA cleaving enzymes created by combining DNA-binding motifs (transcription activator-like effectors or TALES) from plant pathogenic bacteria to a DNA-cutting domain from a nuclease. This method can produce restriction enzymes for any DNA sequence.

transcription factor A DNA-binding protein that binds to specific sequences adjacent to or within the promoter region of a gene; regulates gene transcription.

transcription factory A site of transcription in the nucleus that contains many active RNA polymerase molecules and transcription regulatory molecules. Genes that are regulated by the same transcription factors may be transcribed in the same transcription factory.

transcription Transfer of genetic information from DNA by the synthesis of a complementary RNA molecule using one strand of the DNA as a template.

transcriptional activators Proteins that bind to DNA sequences known as enhancers to increase the rate of gene transcription.

transcriptional repressors Proteins that bind to DNA sequences known as silencers to reduce or block gene transcription.

transcriptome analysis See also transcriptomics. Analysis of all genes expressed by a cell or tissue can include quantitative analysis of gene expression also.

transcriptomics The set of mRNA molecules present in a cell at any given time.

transduction Virally mediated bacterial recombination. Also used to describe the transfer of eukaryotic genes mediated by a retrovirus.

transfer RNA (tRNA) A small ribonucleic acid molecule with an essential role in *translation*. tRNAs contain: (1) a three-base segment (anticodon) that recognizes a codon in mRNA; (2) a binding site for the specific amino acid corresponding to the anticodon; and (3) recognition sites for interaction with ribosomes and with the enzyme that links the tRNA to its specific amino acid.

transformation Heritable change in a cell or an organism brought about by exogenous DNA. Known to occur naturally and also used in *recombinant DNA* studies.

transgene A DNA sequence, often containing a gene or part of a gene, which is isolated from one organism and introduced into a different organism or cells of a different organism.

transgenic organism An organism whose genome has been modified by the introduction of external DNA sequences into the germ line.

translation The derivation of the amino acid sequence of a polypeptide from the base sequence of an mRNA molecule in association with a ribosome and tRNAs.

translational medicine Also called translational genetics. Refers to moving research discoveries in genetics and other disciplines from the laboratory bench to the bedside to improve human health by disease prevention and the treatment of diseases.

translocation A chromosomal mutation associated with the reciprocal or nonreciprocal transfer of a chromosomal segment from one chromosome to another. Also denotes the movement of mRNA through the ribosome during translation.

transmission genetics The field of genetics concerned with heredity and the mechanisms by which genes are transferred from parent to offspring.

transposable element A DNA segment that moves to other sites in the genome, essentially independent of sequence homology. Usually, such elements are flanked at each end by short inverted repeats of 20–40 base pairs. Insertion into a structural gene can produce a mutant phenotype. Insertion and excision of transposable elements depend on two enzymes, transposase and resolvase. Such elements have been identified in both prokaryotes and eukaryotes.

transversion A base change in a DNA molecule that substitutes a purine for a pyrimidine or vice versa.

trinucleotide repeat A tandemly repeated cluster of three nucleotides (such as CTG) within or near a gene. Certain diseases (myotonic dystrophy, Huntington disease) are caused by expansion in copy number of such repeats.

trisomy The condition in which a cell or an organism possesses two copies of each chromosome except for one, which is present in three copies (designated $2n + 1$).

tubulin Protein making up microtubules characteristic of mitosis and meiosis.

tumor-suppressor gene A gene that encodes a product that normally functions to suppress cell division. Mutations in tumor-suppressor genes result in the activation of cell division and tumor formation.

tumorigenesis The development of a tumor.

two-dimensional gel electrophoresis Biochemistry technique for separating peptides or proteins by mass and charge.

ubiquitin ligase A eukaryotic enzyme that catalyzes the addition of a ubiquitin molecule to specific target proteins to tag them for degradation by the proteasome.

ubiquitin A small eukaryotic protein that when covalently attached to other proteins, tags them for degradation by the proteasome.

ubiquitination The process of adding a unit of ubiquitin to a target protein, which serves as a tag for degradation of that target protein by the proteasome.

unit factors The term used by Mendel to describe hereditary factors controlling specific traits. In modern terms, alleles of a specific gene.

variable gene activity hypothesis The idea that in development, differentiation is accomplished by the selective activation and inactivation of genes at different times and in different cell types of the organism.

variable number tandem repeats (VNTRs) Short, repeated DNA sequences (of 2–20 nucleotides) present as tandem repeats between two restriction enzyme sites. Variation in the number of repeats creates DNA fragments of differing lengths following restriction enzyme digestion. Used in early versions of *DNA fingerprinting*.

variance (s^2) A statistical measure of the variation of values from a central value, calculated as the square of the standard deviation.

vector In recombinant DNA, an agent such as a phage or plasmid into which a foreign DNA segment will be inserted and used to transform host cells.

vertical gene transfer The transfer of genetic information from parents to offspring generation after generation.

virulent phage A bacteriophage that infects, replicates within, and lyses bacterial cells, releasing new phage particles.

Western blotting An analytical technique in which proteins are separated by gel electrophoresis and transferred by capillary action to a nylon membrane or nitrocellulose sheet. A specific protein can be identified through hybridization to a labeled antibody.

whole-genome sequencing High-throughput techniques for sequencing all of the DNA in a genome and organizing sequencing data to produce a complete genome sequence.

whole-genome shotgun cloning See *shotgun cloning*.

wobble hypothesis An idea proposed by Francis Crick, stating that the third base in an anticodon in tRNA that can align in several ways to allow it to recognize more than one base in the codons of mRNA.

X-inactivation center (*Xic*) Region of the p-arm of the X-chromosomes in humans which produces a non-coding RNA (*Xist*) that is essential for X-inactivation mechanisms.

X-inactive specific transcript (*XIST*) A locus in the X chromosome inactivation center that controls inactivation of the X chromosome in mammalian females.

X-linkage The pattern of inheritance resulting from genes located on the X chromosome.

XRN1 A eukaryotic exoribonuclease involved in mRNA degradation.

Y chromosome The sex chromosome in species where the male is heterogametic (XY).

yeast artificial chromosomes (YACs) A cloning vector in the form of a yeast artificial chromosome, constructed using chromosomal components including telomeres (from a ciliate) and centromeres, origin of replication, and marker genes from yeast. YACs are used to clone long stretches of eukaryotic DNA.

zinc-finger motif A class of DNA-binding domains seen in proteins. They have a characteristic pattern of cysteine and histidine residues that complex with zinc ions, throwing intermediate amino acid residues into a series of loops or fingers.

zinc-finger nucleases (ZFNs) Sequence-specific DNA cleaving enzymes consisting of a zinc-finger motif (a cluster of two cysteine and two histidine residues that bind zinc atoms). ZFNs provide a mechanism for modifying sequences in the genome in a sequence-specific targeted way.

zip code Sequences found in many mRNAs that serve as binding sites for RNA-binding proteins that influence transcript localization within the cell and translational control.

zip code binding protein 1 (ZBP1) One of a family of highly conserved RNA-binding proteins that play important roles in the localization, stability, and translational control of mRNAs.

zygote The diploid cell produced by the fusion of haploid gametic nuclei.

zygotic genes Gene sets that are activated and expressed in the zygote and early stages of development. In *Drosophila*, these include segmentation genes and homeotic selector genes.

ZZ/ZW Sex-determining chromosomes in animals such as chickens, frogs, and certain fish, where the maleness is determined by homomorphic chromosomes (ZZ) and females by heteromorphic chromosomes (ZW).

Credits

PHOTO CREDITS

About the Authors Courtesy of William S. Klug, Michael R. Cummings, Charlotte A. Spencer, Michael A. Palladino, and Darrell J. Killian.

Chapter 1: CO1.1 Sinclair Stammers/Science Source; CO1.2 Mark Smith/Science Source; CO1.3 Darwin Dale/Science Source; F1.1 Wellcome Library, London; F1.2 Biophoto Associates/Science Source; F1.3 Biophoto Associates/Science Source; F1.4 Dr. Alexey Khodjakov/Science Source; F1.6.1 Photo Researchers/Science Source; F1.6.2 Photo Researchers/Science Source; F1.10 Eye of Science/Science Source; F1.11 Photo courtesy of The Roslin Institute, The University of Edinburgh, Roslin, Scotland, UK; F1.13a Redmond O. Durrell/Alamy Stock Photo; F1.13b Hermann Eisenbeiss/Science Source; F1.14a Dr. Jeremy Burgess/Science Source; F1.14b David McCarthy/Science Source.

Chapter 2 CO2 Heneen Waheeb; F2.2 CNRI/Science Source; F2.4 Leonard Lessin/Science Source; F2.7 (a – f) Waheeb Heneen; F2.14a Biophoto Associates/Science Source; F2.14b Andrew Syred/Science Source; F2.14c Biophoto Association/Science Source.

Chapter 3 CO3 National Library of Medicine; UNF3.1.1 Elenamiv/Shutterstock; UNF3.1.2 Hellens, R. P., et al. Identification of Mendel's White Flower Character. PLoS One 5(10): e13230 (2010). doi:10.1371/journal.pone.0013230; F3.14 Martin Shields/Alamy Stock Photo; UNF3.2 Jenny Dean/REX/Shutterstock.

Chapter 4 CO4 Juniors Bildarchiv GmbH/Alamy Stock Photo; F4.1 John Kaprielian/Science Source; F4.4.1 Jackson Laboratory; F4.9 RoJo Images/Shutterstock; F4.12.1 Photo Researchers/Science Source; F4.14.1 Prisma by Dukas Presseagentur GmbH/Alamy Stock Photo; F4.15 Volodymyr Martyniuk/Shutterstock; F4.16 Stanislav Fridkin/Shutterstock; F4.17 Joel Eissenberg; F4.18ab Joel Eissenberg; F4.19a P. Wegner/Arco Images/AGE Fotostock; F4.19b Dr. William S. Klug; UNF4.1 Ralph Somes; UNF4.2 Ralph Somes; UNF4.3 Ralph Somes; UNF4.4 Ralph Somes; UNF4.5 Shout It Out Design/Shutterstock; UNF4.6 Zuzule/Shutterstock; UNF4.7 Julia Remezova/Shutterstock.

Chapter 5 CO5 Stanley Sessions; F5.16 Arturo Londono, ISM/Science Photo Library.

Chapter 6 CO6 Dr. L. Caro/Science Source; F6.2 Pearson Education; F6.11a SPL/Science Source; F6.13 M. Wurtz/Biozentrum, University of Basel/Science Source; F6.15 Christine Case; F6.18 Pearson Education.

Chapter 7 CO7 Wessex Reg. Genetics Centre, Wellcome Images; F7.2 Catherine G. Palmer; F7.3a Biophoto Associates/Science Source; F7.4 Michael Abbey/Science Source; F7.6a Sari Oneal/Shutterstock; F7.6b Dr. William S. Klug; UNF7.1 Texas A&M University/Newscom; F7.9a Dr. Maria Gallegos, Associate Professor at California State University, East Bay.

Chapter 8 CO8 NHI/National Human Genome Research Institute; F8.2 Arco Images GmbH/Alamy Stock Photo; F8.3 (left) Illustration from Genetic Counseling Aids, 3rd Edition, Copyright 1995, Permission for use granted by Greenwood Genetic Center.; F8.3 (right) Design Pics Inc/Alamy Stock Photo; UNF7.1 Pearson Education; F8.5 Pr Philippe Vago, ISM/Science Photo Library; F8.9 National Cotton Council of America; F8.12 (left) L. Willatt, East Anglian Regional Genetics Service/Science Source; F8.12 (right) Cri Du Chat Support Group; F8.14 Mary Lilly/The Observatories for the Carnegie Institution for Science; F8.19 Christine Harrison.

Chapter 9 CO9 Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. Kukat C, et.a;l. Proc Natl Acad Sci U S A. 2011 Aug 16;108(33):13534—9.; F9.1 Tikta Alik/Shutterstock; F9.2 Dartmouth College Electron Microscope Facility; F9.3 Montenegro-Montero A, Goity A, Larrondo LF (2015) The bZIP Transcription Factor HAC-1 Is Involved in the Unfolded Protein Response and Is Necessary for Growth on Cellulose in Neurospora crassa. PLoS ONE10(7): e0131415. https://doi.org/10.1371/journal.pone.0131415; F9.4 W. C. Copeland et. al. (2003). POS5 Gene of Saccharomyces cerevisiae Encodes a Mitochondrial NADH Kinase Required for Stability of Mitochondrial DNA. Eukaryotic Cell, 2(4):809—820.; F9.6 Dr. Richard Kolodner; F9.7 Don W. Fawcett/Science Source; F9.9 (a, b) Abu-Amero KK, et.al. (2009). A patient with typical clinical features of mitochondrial encephalopathy, lactic acidosis and stroke-like episodes (MELAS) but without an obvious genetic cause: a case report. Journal of Medical Case Reports, 3:77; F9.10 Koteshwar Rao/Flickr Open/Getty Images; F9.11 Courtesy of Dr. Reiko Kuroda; F9.12 Uwe Irion.

Chapter 10 CO10 Richard Megna/Fundamental Photographs, NYC; F10.4 Meckes/Ottawa/Science Source; F10.11 Arizona Board of Regents/ASU Ask A Biologist F10.17 Ventana Medical Systems, Inc.; F10.18 Permission Granted by John Wiley and Sons, from Gel Electrophoresis by R. Westermeier, Serva Electrophoresis GmbH, Heidelberg, Germany, ELS (February 2013).

Chapter 11 CO11 Dr Gopal Murti/Science Source; F11.5 J.H. Taylor, (1963). The Replication and Organization of DNA in Chromosomes, Molecular Genetics, Part 1, pp. 74 and 75, Academic Press, New York, 1963.; UNF11.1 Republished with permission of [American Society for Microbiology], from Early Embryonic Lethality Due to Targeted Inactivation of DNA Ligase III by Puebla-Osorio N., et al,. Mol Cell Biol. 26;10 3935–3941, (2006 May); permission conveyed through Copyright Clearance Center, Inc.; F11.14 Diffley, J.FX., (2016). Louis-Jeantet Prize Winner: Perspective. On the road to replication. EMBO Molecular Medicine, 9(11): 1463–1621.; 253: Victoria Foe; F11.18 David Dressler and Huntington Potter.

Chapter 12 CO12 Don. W. Fawcett/Science Source; F12.2 Biology Pics/Science Source; F12.1a Dr. Michel Wurtz/Biozentrum, University of Basel/Science Source; F12.1b Dr. William S. Klug; F12.3 G. Murti/Science Source; F12.5 Republished with permission of The Company of Biologists, Ltd., Identification of the Drosophila interband-specific protein Z4 as a DNA-binding zinc-finger protein determining chromosomal structure by Harald Eggert, et. al., Journal of Cell Science 17: 4253–4264 2004; permission conveyed through Copyright Clearance Center, Inc.; F12.6 John Ellison; F12.7a Science Source; F12.7b Dr. William S. Klug; F12.8 Victoria Foe; F12.11 UVM Genetics & Genomics Wiki; F12.12 Cultura RM/Alamy Stock Photo; F12.16 Dr. William S. Klug.

Chapter 13 CO13 Professor Oscar L. Miller/Science Source; 305: Nobel Media Ab 2017; F13.16 From Miller, O. L. et al. Visualization of bacterial genes in action. Science 169:392–395, (1970). Reprinted with permission from AAAS.

Chapter 14 CO14 From the front cover of Science, Vol. 292, Issue 5518, pp. 870–872, May 4, 2001.; F14.9a Copyright holder is Cold Spring Harbor Laboratory Press; F14.9b Elena Kiseleva, ICiG, Russia; F14.14 Sebastian Kaulitzki/Shutterstock; F14.19 Ken Eward/Science Source.

Chapter 15 CO15 Gerry Neuffer; F15.15 Hadj/Sipa/Newscom; UNF15.1 (a,b) From Starr, T. K. et al. A Transposon-Based Genetic Screen in Mice Identifies Genes Altered in Colorectal Cancer. Science 323(5922):1747–1750, (2009). Reprinted with permission from AAAS.

Chapter 17 CO17 Bolzer A., et.al. (2005). Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLoS Biology, 3(5): e157.; F17.2 Osborne, C.S. et al. (2007) Myc Dynamically and Preferentially Relocates to a Transcription Factory Occupied by Igh. PLoS Biology, 5(8): e192.

Chapter 18 CO18 Pearson Education; F18.4 Pearson Education; F18.6 Pearson Education; UNF18.1 Hasuwa, H. et.al. (2013). MiR-200b and miR-429 Function in Mouse Ovulation and Are Essential for Female Fertility. Science, 341(6141):71–73; F18.8 Pearson Education; F18.9 Pearson Education; F18.11 Hochstrasser M. (2009). Origin and function of ubiquitin-like proteins. Nature, 458(7237):422–9.

Chapter 19 CO19 Courtesy of the John Innes Centre, Norwich.; F19.13 Dolinoy et al. (2006). Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environmental Health Perspectives, 114(4):567–72.

Chapter 20 CO20 Pascal Goetgheluck/Science Source; F20.3a SPL/Science Source; F20.5b Reproduced with permission from Sigma-Aldrich Co. LLC; F20.9 Scott Sinklier/Alamy Stock Photo; F20.11 Proceedings of the National Academy of Sciences; F20.12 From Wan, L., Dockendorff, et al. Characterization of dFMR1, a Drosophila melanogaster Homolog of the Fragile X Mental Retardation Protein. Molecular and Cellular Biology, 20(22):8536–8547, (2000). doi: 10.1128/MCB.20.22.8536–8547.2000. Reproduced/amended with permission from American Society for Microbiology.; F20.13 Gerald B. Downes; F20.18a RAGUET/PHANIE/AGE Fotostock; F20.18b Michael A. Palladino; F20.20a Eye of Science/Science Source; F20.20b Photograph courtesy of Ralph Brinster. University of Pennsylvania School of Veterinary Medicine.

Chapter 21 F21.11 Pawel Stankiewicz;; F21.14.1 Volker Steger/Science Source; F21.14.2 dra_schwartz/E+/Getty Images; F21.15 TEK IMAGE/Science Photo Library RF/AGE Fotostock.

Chapter 22 CO22 Harris Wang; F22.1a Eric Carr/Alamy Stock Photo; F22.2 Courtesy of John Doebley; F22.4 CB2/ZOB/WENN/Newscom; F22.9 Affymetrix/Thermo Fisher Scientific; F22.11a Sebastian Kaultizki/Shutterstock; F22.11b CDC/C. D. Humphrey and T. G. Ksiazek; F22.11c CDC/National Escherichia, Shigella, Vibrio Reference Unit at CDC; F22.11d Centers for Disease Control and Prevention; F22.15 Courtesy of University of California, San Diego.

Chapter 23 CO23 Edward B. Lewis, California Institute of Technology; F23.1a Steve Gschmeissner/Science Source; F23.1b Georg Halder; F23.9 (a, b) Peter Lawrence; F23.10 Paddock, Stephen; F23.11 Karagüzel G, Aktürk FA, Okur E, Gümele HR, Gedik Y, Ökten A. Cleidocranial Dysplasia: A Case Report. Journal of Clinical Research in Pediatric Endocrinology. 2010;2(3):134–136. doi:10.4274/jcrpe.v2i3.134.; F23.12a Republished with permission of Elsevier, from [Cbfa1, a Candidate Gene for Cleidocranial Dysplasia Syndrome, Is Essential for Osteoblast Differentiation and Bone Development. Otto F., et al. 89;5 765–771, (May 1997)]; permission conveyed through Copyright Clearance Center, Inc.; F23.13a Eye of Science/Science Source; F23.17a MPI for Developmental Biology/Dept. for Molecular Biology; F23.19 (a, b) Dr. Elliot M. Meyerowitz; UNF23.1 (a, b) Reprinted by permission from Macmillan Publishers Ltd: Reactivating head regrowth in a regeneration-deficient planarian species. Liu S.Y, et al, Nature Reviews Genetics, 500;7460, 81–84 (August 2013).; F23.24b Gehring WJ. Chance and Necessity in Eye Evolution. Genome Biology and Evolution. 2011;3:1053–1066. doi:10.1093/gbe/evr061.; F23.24 Jurgen Berger and Ralf Dahm, Max Plank Institute for Developmental Biology, Tubingen, Germany; F23.25a Georg Halder; F23.26a Courtesy of Univeristy of Basel, Biozentrum; F23.25b Courtesy of University of Basel, Biozentrum; T23.5 Pearson Education.

Chapter 24 CO24 SPL/Science Source; F24.1(a,b) Hesed M. Padilla-Nash/Antonio Fargiano/National Institute of Health Genetics; F24.3 (a,b) Manfred Schwab; F24.8 Dr. Gopal Murti/Science Source.

Chapter 25 CO25 2009fotofriends/Shutterstock; F25.9 Rekemp/iStock/Getty Images; UNF 25.1 Nature's Images/Science Source.

Chapter 26 CO26 Simon Booth/Science Source; F26.1 Pixshots/Shutterstock; F26.15 Republished with permission of [Elsevier]; from A 122.5-Kilobase Deletion of the P Gene Underlies the High Prevalence of Oculocutaneous Albinism Type 2 in the Navajo Population. Yi Z, et. al., American Journal of Human Genetics. 72(1):62–72, 2003; permission conveyed through Copyright Clearance Center, Inc.; F26.18 Smithsonian Institution Libraries; F26.20a Niels Poulsen DK/Alamy Stock Photo; F26.20b Niklas Lindqvist/National Ciklid Society; F26.22a Mchugh Tom/Science Source/Getty Images; F26.22b: Peter Scoones/The Image Bank/Getty Images; F26.25 RIA Novosti/Science Source.

ST4 STF4.1 Pascal Pavani/AFP/Getty Images; STUNF4.1 AquaBounty Technologies; STF4.3Courtesy of International Rice Research Institute; STF4.6 $©$ 2018 Courtesy of Bio-Rad Laboratories, Inc.; STUNF4.1 Kheng Guan Toh/Shutterstock.

ST6 STUNF6.1 Acey Harper/The LIFE Images Collection/Getty Images.

TEXT CREDITS

Chapter 5: Page 98 Sturtevant, A.H. (1965), A History of Genetics, Harper & Row.

Chapter 6 T6.2 Pearson Education.

Chapter 7 T7.1 Pearson Education.

Chapter 8 Page 191 Francis Galton (1904). Eugenics: Its Definition, Scope, and Aims. American Journal of Sociology, 10:1, 1–25.

Chapter 10 F10.2 Avery, O.T., et.al. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. The Journal of Experimental Medicine, 79(2):137–158.; T10.3 Pearson Education; F10.14 ; 228: J. D. Watson, F. Crick (1953). A Structure for Deoxyribose Nucleic Acid. Nature, 421(6921): 397–8.; 231: Ventana Medical Systems, Inc.

Chapter 11 Pearson Education.

Chapter 13 T13.1 Source: After M. Nirenberg and J. H. Matthaei (1961).

Chapter 14 F14.10 Pearson Education; F14.11 AKU Society.

Chapter 18 F18.2 Reprinted by permission from Macmillan Publishers Ltd: Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics by Keren, H. et. al., 11: 345–355 (May 2010).; Chapter 19 T19.1 Lokk K., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biology, 15(4):r54. Table 2.

Chapter 19 F19.3 Pearson Education; F19.6a Courtesy of Eric di Luccio; F19.6b Bannister, A.L., Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research, 21:381–395.; F19.7 (a – d) Reproduced with permission of Annual Review of, Annual Review of Biochemistry Volume 81 $©$ July 2012 by Annual Reviews, http://www.annualreviews.org.; F19.8 Reproduced with permission of Annual Review of, Annual Review of Biochemistry Volume 81 © July 2012 by Annual Reviews, http://www.annualreviews.org.; F19.9 Adapted by permission from Macmillan Publishers Ltd: Rosanna Weksberg, Cheryl Shuman, and J Bruce Beckwith. Beckwith–Wiedemann syndrome, European Journal of Human Genetics (2010) 18, 8–14, p. 10, fig.2; F19.10 Adapted from Eckersley-Maslin MA, Thybert D, et.al. (2014). Random Monoallelic Gene Expression Increases upon Embryonic Stem Cell Differentiation. Developmental cell, 28(4):351–365. doi:10.1016/j.devcel.2014.01.017.; F19.11 Banno, Kouji, et al. (2013. Epimutation in DNA Mismatch Repair (MMR) Genes. Biochemistry, Genetics and Molecular Biology “New Research Directions in DNA Repair”, book edited by Clark Chen, May 22, 2013 under CC BY 3.0 license. https://www.intechopen.com/books/new-research-directions-in-dna-repair/epimutation-in-dna-mismatch-repair-mmr-genes; F19.14 Reproduced with permission from the Annual Review of Genomics and Human Genetics, Volume 9 © 2008 by Annual Reviews, http://www.annualreviews.org.

Chapter 20 F20.15 Pearson Education; F20.16 Pearson Education; F20.19 Pearson Education.

Chapter 21 T21.2 Originally adapted from Palladino, M. A. Understanding the Human Genome Project, 2nd ed. Benjamin Cummings, 2006.; Republished with permission of AAAS, from Stanley Fields, Molecular Biology. Site-seeing by sequencing. Science 316(5830):1441-1442, June 2007; permission conveyed through Copyright Clearance Center, Inc.; F21.10 Reprinted by permission from Macmillan Publishers Ltd: [Nature] News Feature, Human Genome at ten: The sequence explosion. Nature 464:670- 671, (2010 March).; F21.12 Republished with permission of [Oxford University Press], from [SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips Marcus, S. Lee, H., Schatz, M.C., vol.30(24):3476–3483 (2014)]; permission conveyed through Copyright Clearance Center, Inc.; F21.13 Qin, N., Yang, F., et al. Alterations of the human gut microbiome in liver cirrhosis. Nature, 513(7516):59–64. F21.16.2 Republished with permission of [ELECTROPHORESIS. SWISS-2DPAGE: A database of two-dimensional gel electrophoresis images. Appel, Ron D. et.al. vol. 14(1) pp. 1232–1238 (1993)] ; permission conveyed through Copyright Clearance Center, Inc.; F21.18 Republished with permission of [American Association for the Advancement of Science (AAAS)], from [Protein Sequences from Mastodon and Tyrannosaurus Rex Revealed by Mass Spectrometry. Asara J., et al., Science 316(5822):280–285, 2007 April]; permission conveyed through Copyright Clearance Center, Inc.

Chapter 22 F22.10 Reprinted by permission from Macmillan Publishers Ltd: [Distinct types of diffuse large B-cell lymphoma identified by gene expression profling. Ash A. Alizadeh et. al.] Nature 403, 503–511, (February 2000); F22.12a Reprinted by permission from Macmillan Publishers Ltd: [Single-cell RNA-seq identifies a PD-1hi ILC progenitor and defines its development pathway]. Yu et al., Nature 539, 102–106, (2016).; F22.13 Jones, G. T., et.al. (2017). Meta-Analysis of Genome-Wide Association Studies for Abdominal Aortic Aneurysm Identifies Four New Disease-Specific Risk Loci, Circulation Research. 120(2):341–353.; F22.14b Reprinted with permission from The Science and Applications of Synthetic and Systems Biology: Workshop Summary, 2011 by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.; UNF22.1 Rick Weiss., June 29, 2007. Scientists Report DNA Transplant Organisms Adopt Donor Traits. The Washington Post; UNF22.1 Reprinted by permission from Macmillan Publishers Ltd: [Nature] Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. Petukhova et al., 466, 113–117, (2010).; UNF22.2 Modified from Lee et al. 2001. Infect. and Immunity 69: 5786–5793.

Chapter 23 F23.2 Pearson Education; F23.5a From Evolution & Development Gehring, The animal body plan, the prototypic body segment, and eye evolution. Walter. J. Sept. 2012, 14(1)34–46. Reproduced with permission of John Wiley & Sons, Inc.; F23.7 Paddock, Stephen.; T23.4 From Gerhart, J. (1999). Warkany lecture: Signaling pathways in development. Teratology, 60(4):226-39. Reproduced with permission of John Wiley & Sons, Inc., F23.26a Courtesy of University of Basel, Biozentrum.

Chapter 24 F24.2 Elizabeth Cook; UNT24.1 Pearson Education.

Chapter 25 F25.8 Reprinted by permission from Macmillan Publishers Ltd: Quantitative trait loci in Drosophila by Mackay, TF. Nature Reviews Genetics, 2:11–20 (January 2001).; T25.5 Pearson Education.; F25.10 Republished with permission of [American Society for Microbiology], from Increase in tomato locule number is controlled by two single nucleotide polymorphisms located near WUSCHEL. Plant Physiology, by Munos, S. et al., 156(4):2244–2254, August 2011; permission conveyed through Copyright Clearance Center, Inc.; F25.11: Republished with permission of American Association for the Advancement of Science, from RNA splicing is a primary link between genetic variation and disease. Science 352(6285):600–604, Li, Y., et al., (April 2016); permission conveyed through Copyright Clearance Center, Inc.; UNF25.1 Nature's Images/Science Source.

Chapter 26 F26.2 Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature, 304(5925):412–417.; F26.6 Reprinted by permission from Macmillan Publishers Ltd: Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR- 5 chemokine receptor gene by Samson M., et al., Nature 382(6593):725 (August 1996).; F26.24 Reproduced with permission from the Annual Review of Genomics and Human Genetics, Volume 11 $©$ 2010 by Annual Reviews, http://www.annualreviews.org; F26.25 RIA Novosti/Science Source.

ST1 STF1.2 Pearson Education; STF1.4 Pearson Education; STF1.5 Pearson Education.

ST2 STT2.2 Reprinted with permission from the Journal of Forensic Sciences, Vol 48, Issue 4, Copyright ASTM International, 100 Barr Harbor Drive, West Conshohocken, PA 19428; STF2.2 Reproduced with permission from Promega Corporation; STF2.4 Reproduced with permission from Promega Corporation; STF2.7 Parabon NanoLabs, Inc.

ST3 ST3.1a Reproduced with permission of Dako Denmark A/S, a subsidiary of Agilent Technologies, Inc., Santa Clara, California, USA. All rights reserved; STF3.1b Reproduced with permission of Dako Denmark A/S, a subsidiary of Agilent Technologies, Inc., Santa Clara, California, USA. All rights reserved; STF3.2 Pearson Education; STF3.4 Pearson Education.

ST4 STT4.1 Pearson Education.

ST5 STF5.1 Pearson Education; STF5.2 Pearson Education; STF5.3 Targeting DNA. After 20 years of high-profile failure, gene therapy is finally well on its way to clinical approval. Jef Akst. Reprinted with permission of The Scientist from “Targeting DNA” July, 2002; STF5.8 Pearson Education.

ST6 Page 710 ”On Chorea,” by George Huntington, M.D., originally appeared in The Medical and Surgical Reporter: A Weekly Journal (Philadelphia: S.W. Butler), 26(15):317–321, (April 1872); page 711 “On Chorea,” by George Huntington, M.D., originally appeared in The Medical and Surgical Reporter: A Weekly Journal (Philadelphia: S.W. Butler), 26(15):317–321, (April 1872); STT6.1 From The Huntington's Disease Collaborative Research Group (1993). Cell 72: 971–983, Table 3, p. 976 STF6.1 Pearson Education; STF6.3 Prion-like Behavior in the Huntingtin Protein by Ocords, Protein Aggregation (May 2015); STF6.4 Rick Morimoto/Morimoto Lab; STF6.6 Permission granted by PNAS from Synthetic zinc finger repressors reduce mutant huntingtin expression in the brain of R6/2 mice by Garriga-Canut M, et.al. 109(45):E3136–45. (Oct. 2012); STF6.7 Pearson Education; STF6.8 Jong-Min Lee; STF6.9 Pearson Education.

Index

Note: **Bold** page numbers indicate key terms. Page numbers followed by b, f, or t indicate boxes, figures, or tables respectively.

EVOLVING CONCEPT OF THE GENE

The Evolving Concept of the Gene is a unique feature, integrated in key chapters, that highlights how scientists' understanding of the gene has changed over time. By underscoring how the conceptualization of the gene has evolved, our goal is to help students appreciate the process of discovery that has led to an ever more sophisticated understanding of hereditary information.

CHAPTER 3 **pg. 48** Based on the pioneering work of Gregor Mendel, the gene was viewed as a heritable unit factor that determines the expression of an observable trait, or phenotype. ■

CHAPTER 4 **pg. 70** Based on the work of many geneticists following the rediscovery of Mendel's work in the very early part of the twentieth century, the chromosome theory of inheritance was put forward, which hypothesized that chromosomes are the carriers of genes and that meiosis is the physical basis of Mendel's postulates. In the ensuing 40 years, the concept of a gene evolved to reflect that this hereditary unit can exist in multiple forms, or alleles, each of which can impact on the phenotype in different ways, leading to incomplete dominance, codominance, and even lethality. It became clear that the process of mutation was the source of new alleles. ■

CHAPTER 5 **pg. 110** Based on the gene-mapping studies in *Drosophila* and many other organisms from the 1920s through the mid-1950s, geneticists regarded genes as hereditary units organized in a specific sequence along chromosomes, between which recombination could occur. Genes were thus viewed as indivisible "beads on a string." ■

CHAPTER 6 **pg. 144** In the early 1950s, genes were regarded as indivisible units of heredity that could undergo mutation and between which *intergenic* recombination could occur. Seymour Benzer's pioneering genetic analysis of the *rII* locus in phage T4 in the mid-1950s demonstrated that the gene is not indivisible. Instead, he established that multiple sites exist within a gene, each capable of undergoing mutation, and between which *intragenic* recombination can occur. Benzer was able to map these sites within the *rII* locus. As we will see in an ensuing chapter, around the same time, it became clear that the gene is composed of DNA nucleotide pairs, each of which can be the site of mutation or recombination. ■

CHAPTER 10 **pg. 229** Based on the model of DNA put forward by Watson and Crick in 1953, the gene was viewed for the first time in molecular terms as a sequence of nucleotides in a DNA helix that encodes genetic information. ■

CHAPTER 13 **pg. 304** The elucidation of the genetic code in the 1960s supported the concept that the gene is composed of an uninterrupted linear series of triplet nucleotides encoding the amino acid sequence of a protein. While this is indeed the case in bacteria and viruses that infect them, in 1977, it became apparent that in eukaryotes, the gene is divided into coding sequences, called exons, which are interrupted by noncoding sequences, called introns, which must be spliced out during production of the mature mRNA. ■

CHAPTER 14 **pg. 328** In the 1940s, a time when the molecular nature of the gene had yet to be defined, groundbreaking work by Beadle and Tatum provided the first experimental evidence concerning the product of genes, their "one-gene:one-enzyme" hypothesis. This idea received further support and was later modified to indicate that one gene specifies one polypeptide chain or functional RNA. ■

CHAPTER 16 **pg. 383** The groundbreaking work of Jacob, Monod, and Lwoff in the early 1960s, which established the operon model for the regulation of gene expression in bacteria, expanded the concept of the gene to include noncoding regulatory sequences that are present upstream (5′) from the coding region. In bacterial operons, the transcription of several contiguous structural genes whose products are involved in the same biochemical pathway is regulated in a coordinated fashion. ■

CHAPTER 17 **pg. 407** Based on the findings of the ENCODE project, we now know that DNA sequences that have previously been described as "junk" DNA, since they do not encode polypeptides, are nonetheless often transcribed into what we call noncoding RNA (ncRNA). Since the function of some of these RNAs is now being determined, we must consider whether the concept of the gene should be expanded to include these DNA sequences. Currently, this is being debated, and given that this is the last "Evolving Concept of the Gene" entry in the text, we are wondering what you think, based on your study of genetics? ■