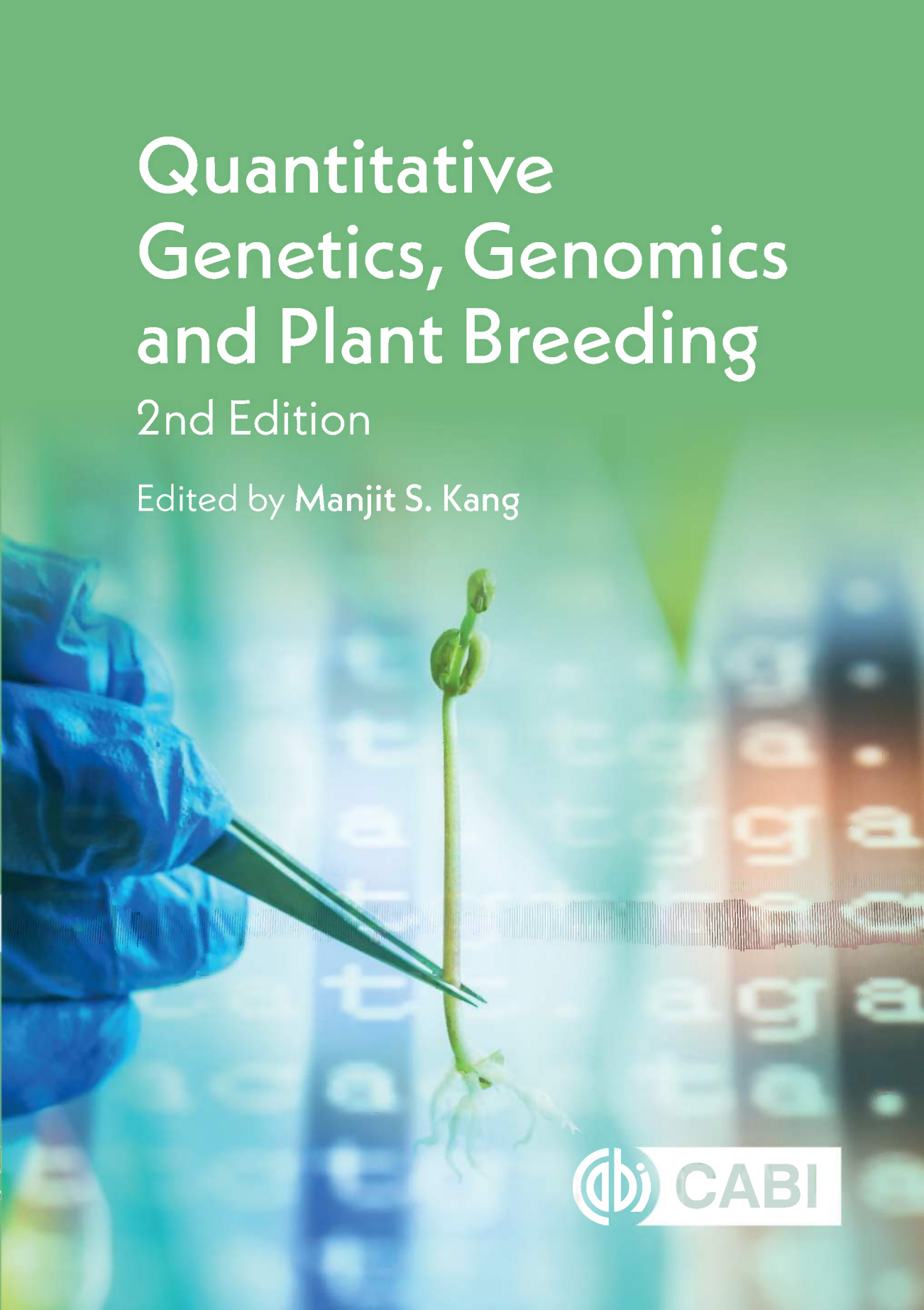
Quantitative Genetics, Genomics and Plant Breeding

2nd Edition

Edited by Manjit S. Kang



CABI

Quantitative Genetics, Genomics and Plant Breeding

Second Edition

Quantitative Genetics, Genomics and Plant Breeding

Second Edition

Edited by

Manjit S. Kang

Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

CABI

CABI is a trading name of CAB International

CABI
Nosworthy Way
Wallingford
Oxfordshire OX10 8DE
UK

Tel: +44 (0)1491 832111
Fax: +44 (0)1491 833508
E-mail: info@cabi.org
Website: www.cabi.org

CABI
745 Atlantic Avenue
8th Floor
Boston, MA 02111
USA

Tel: +1 (617)682-9015
E-mail: cabi-nao@cabi.org

© CAB International 2020. All rights reserved. No part of this publication may be reproduced in any form or by any means, electronically, mechanically, by photocopying, recording or otherwise, without the prior permission of the copyright owners.

A catalogue record for this book is available from the British Library, London, UK.

Library of Congress Cataloging-in-Publication Data

Names: Kang, Manjit S., editor.
Title: Quantitative genetics, genomics and plant breeding / Manjit S. Kang, Adjunct Professor (Quantitative Geneticist), Department of Plant Pathology, Kansas State University, Manhattan, Kansas, USA.
Description: 2nd edition. | Wallingford, Oxfordshire ; Boston : CABI, [2020] | Includes bibliographical references and index. | Summary: "Quantitative genetics and breeding have made a major contribution to crop improvement over the years, but the genomics revolution has dramatically changed the field – this second edition of a popular book explains how traditional and genomic techniques have been combined to advance the field"– Provided by publisher.
Identifiers: LCCN 2019047038 (print) | LCCN 2019047039 (ebook) | ISBN 9781789240214 (hardback) | ISBN 9781789242942 (ebk) | ISBN 9781789242959 (epub)
Subjects: LCSH: Crops–Genetics. | Plant breeding. | Quantitative genetics.
Classification: LCC SB123 .Q36 2020 (print) | LCC SB123 (ebook) | DDC 631.5/3–dc23
LC record available at https://lccn.loc.gov/2019047038
LC ebook record available at https://lccn.loc.gov/2019047039

References to Internet websites (URLs) were accurate at the time of writing.

ISBN-13: 9781789240214 (hardback)
9781789242942 (ePDF)
9781789242959 (ePub)

Commissioning Editor: David Hemming
Editorial Assistant: Emma McCann
Production Editor: Marta Patiño

Typeset by SPi, Pondicherry, India
Printed and bound in the UK by Severn, Gloucester

Dr George P. Rédei, known among his colleagues and students as the 'Encyclopedia of Genetics', passed away on 10 November 2008. This edition of *Quantitative Genetics, Genomics and Plant Breeding* is dedicated to him. His chapter from the first edition of the book is being reproduced here.

Contents

Contributors

N. Ahmadi
CIRAD, UMR AGAP, F-34398, Montpellier, France;
AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
M.S. Anderson
HarvestPlus, c/o International Food Policy Research Institute (IFPRI), Washington DC, USA.
R. Anderson
School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, Australia.
P. Annicchiarico
Research Centre for Animal Production and Aquaculture, viale Piacenza 29, Lodi, Italy.
V.N. Arief
The University of Queensland, Brisbane, Australia.
B. Badu-Apraku
International Institute of Tropical Agriculture, Ibadan, Nigeria.
S.S. Banga
Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India.
J. Bartholomé
CIRAD, UMR AGAP, F-34398, Montpellier, France;
AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
K.E. Basford
The University of Queensland, Brisbane, Australia.
J. Batley
School of Biological Sciences, University of Western Australia, Perth, Australia.
D. Bhatia
Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India.
D.S. Brar
School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India;
Former Head, Plant Breeding, Genetics & Biotechnology Division, International Rice Research Institute, Manila, Philippines.
J. Burgueño
International Maize and Wheat Improvement Center (CIMMYT), México City, México.
T.-V. Cao
CIRAD, UMR AGAP, F-34398, Montpellier, France;
AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.

H. Ceballos
CIAT, Cali, Colombia.
S. Ceccarelli
Consultant, Rete Semi Rurali, via di Casignano 25, Scandicci, Italy.
S. Chapagain
School of Plant, Environmental and Soil Sciences, Louisiana State University Agricultural Center, Baton Rouge, LA, USA.
B. Cherian
HarvestPlus, c/o International Food Policy Research Institute (IFPRI), Washington DC, USA.
J. Crossa
International Maize and Wheat Improvement Center (CIMMYT), México City, México.
M.F. Danilevicz
School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, Australia.
I.H. DeLacy
The University of Queensland, Brisbane, Australia.
D. Edwards
School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, Australia.
M.A.B. Fakorede
Obafemi-Awolowo University, Ile-Ife, Osun State, Nigeria.
C. Tay Fernandez
School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth, Australia.
M. Govindaraj
Crop Improvement, Asia Research Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India.
C. Grenier
CIRAD, UMR AGAP, F-34398, Montpellier, France;
AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
M. Gupta
Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India.
L.A. Hunt
Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada.
S.V. Joshi
South African Sugarcane Research Institute, Durban, South Africa.
P. Juliana
International Maize and Wheat Improvement Center (CIMMYT), México City, México.
A. Kanatti
Crop Improvement, Asia Research Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India.
M.S. Kang
Department of Plant Pathology, Kansas State University, Manhattan, KS, USA.
M.T. Labuschagne
Department of Plant Sciences (Plant Breeding), University of the Free State, Bloemfontein, South Africa.
D. Lloyd Evans
South African Sugarcane Research Institute, Durban, South Africa;
Cambridge Sequence Services (CSS), Waterbeach, Cambridge, UK;
Department of Computer Sciences, Université Cheikh Anta Diop de Dakar, Dakar, Sénégal.
G.S. Miglani
School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India.

S. Mondal
International Maize and Wheat Improvement Center (CIMMYT), México City, México.
O.A. Montesinos-López
Facultad de Telemática, Universidad de Colima, Colima, México.
E. Obeng-Bio
CSIR – Crops Research Institute, Fumesua, Kumasi, Ghana.
S.A. Oyekale
Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria.
P. Pérez-Rodríguez
Colegio de Postgraduados, Montecillos, Edo. de México, México.
W.H. Pfeiffer
HarvestPlus, c/o International Food Policy Research Institute (IFPRI), Washington DC, USA.
M. Pillay
Department of Biotechnology, Vaal University of Technology, Vanderbijlpark, South Africa.
K.N. Rai
Crop Improvement, Asia Research Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India.
L. Singh
School of Plant, Environmental and Soil Sciences, Louisiana State University Agricultural Center, Baton Rouge, LA, USA.
R. Singh
School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India.
R.P. Singh
International Maize and Wheat Improvement Center (CIMMYT), México City, México.
P.K. Subudhi
School of Plant, Environmental and Soil Sciences, Louisiana State University Agricultural Center, Baton Rouge, LA, USA.
A.O. Talabi
International Institute of Tropical Agriculture, Ibadan, Nigeria.
S.N. Tchala
Institut Togolais de Recherche Agronomique (ITRA), Lomé Cacaveli, Togo.
S. Tirnaz
School of Biological Sciences, University of Western Australia, Perth, Australia.
P.S. Virk
HarvestPlus, c/o International Food Policy Research Institute (IFPRI), Washington DC, USA.
W. Yan
960 Carling Ave, Ottawa, Ontario, Canada.

Foreword

The impressive edifice of our modern civilization rests upon stable and sustainable food production. Given that 9 billion people are expected to inhabit the planet in 2050, plant breeders face the enormous (and onerous) task of increasing food production, and not simply in yield but also in nutritional value. They must accomplish this task in the face of climate change, ideally while using fewer resources (acreage, water, fertilizer and pesticides). Further, the majority of this progress must occur in developing nations. Finally, any such genetic improvement must not come at the cost of diminishing genetic diversity, as new challenges for the breeder, such as emerging diseases, are continually arising.

This is the backdrop for the second edition of this collection of reviews on modern methods for plant breeding. As with the initial 2002 edition, it features a strong international cast of researchers, many of whom are focused on plant improvement in developing nations. Much has happened since the 2001 meeting that led to the first edition. Although the word 'genome' appeared in the title of that first book, the first full genome sequence of a crop (rice) was not published until after the 2001 meeting. Since then, not only have many mainstream crops been fully sequenced, there are also efforts to sequence many lesser known ones, such as the African Orphan Crops Consortium's plan to sequence 100 new genomes. Genome editing, essentially unforeseen in 2001, is now an important weapon in the plant breeder's arsenal. During the past two decades, the powerful statistical machinery of mixed-model analysis has moved from its initial home in animal breeding into the toolkit of everyday plant breeders. Genome-wide association studies – very briefly outlined in a paper by Jean-Luc Jannink and myself in the 2002 volume – have also become a standard tool. Importantly, all of these approaches fairly easily transfer to less-developed crops, those in which much of the production gain must be sought.

As with the first volume, Professor Manjit Kang has done a masterful job weaving together a collection of reviews that highlight many of these issues. The task that breeders face is daunting, but given the progress in methods seen during the past two decades, hope remains that they will be up to the task. The key will be in translating these new methods from tools in academic research settings into the accelerated improvement in crops.

Bruce Walsh
University of Arizona, Tucson, Arizona

Preface

The first edition of *Quantitative Genetics, Genomics and Plant Breedin*g – a compilation of 24 chapters authored by prominent scientists from across the world – came out in 2002. The fields of quantitative genetics and genomics, as applied to crop improvement, have been progressing at a rapid pace. It is therefore essential to consolidate the important new information on all pertinent issues relative to the subject matter in the second edition of this title.

Since 2002, climate change effects on food production have become more evident, the role of bioinformatics in crop improvement has become clearer and emphasis on developing climate-resilient crop varieties has increased. In addition, new technologies, such as genome editing for crop improvement, have emerged. Such new, cutting-edge technologies are needed to meet the food demand of the ever-increasing world population, which as of this writing is about 7734 million. The Food and Agriculture Organization of the United Nations estimates that demand for food, feed and fibre will increase by 70% by 2050; the increasing demand is being driven by higher world population, rising incomes and urbanization.

The second edition of *Quantitative Genetics, Genomics and Plant Breedin*g contains 20 authoritative chapters, authored by scientists from global agricultural research institutes, such as CIMMYT, IITA, CIAT, ICRISAT and HarvestPlus; national agricultural research centres and world-class universities.

The first chapter titled 'Vignettes of the History of Genetics' authored by Dr George P. Rédei, a well-known historian of genetics, has been reproduced from the first edition, as this edition is being dedicated to him, as a tribute (Dr Rédei passed away on 10 November 2008). Dr Rédei enumerated historical developments in genetics up to the 21st century. He envisioned the role of quantitative genetics and genomics in crop improvement to increase greatly in the 21st century.

The rest of the book has been divided into two main sections: (i) Quantitative Genetics: Plant Breeding, Bioinformatics, Genome Editing and G×E Interaction (Chapters 2 to 12) and (ii) Intersection of Breeding, Genetics and Genomics: Crop Examples (Chapters 13 to 20). A brief discussion on each chapter follows.

In Chapter 2 (Food and Health: The Role of Plant Breeding), Dr Salvatore Ceccarelli suggests that there is a contradiction between the need for diet diversity and crop uniformity, which is the main feature of industrial agriculture, and that there is also a contradiction between crop uniformity and the need to adapt crops to both short- and long-term climate change (and to associated changes in the spectrum of pests and diseases). The author recommends that plant breeding should shift from 'cultivating uniformity' to 'cultivating diversity' for the benefit of human health.

In Chapter 3 (The Importance of Plant Pan-genomes in Breeding), Dr Tirnaz and colleagues discuss the importance of pan-genomes, indicating that pan-genomics is being used to study

genomic structural variations, including presence–absence variations and copy number variations in plant genomes. They have provided a list of pan-genome studies in plants, including wheat, rice, maize and *Brassica*. The authors suggest that pan-genomes can provide a complete genomic content of a species and add value to all aspects of genomic studies and molecular breeding strategies.

In Chapter 4 (Genome Editing Technologies for Crop Improvement), Dr Michael Pillay discusses the latest genome-editing techniques. He points out that CRISPR/Cas9 technology was first demonstrated in rice, sorghum and wheat, and that, of late, there has been an explosion in research in genome editing of plants using the CRISPR/Cas9 technology and its modifications. Traits improved via genome editing include disease resistance, nutritional improvement and yield of crops. He discusses advantages and shortcomings of the available techniques.

In Chapter 5 (Epigenome Editing in Crop Improvement), Dr Miglani and Mr Singh elaborate on one of the newest techniques on the horizon: 'epigenome editing' (epigenetics = heritable alterations in chromatin architecture that do not involve changes in the underlying DNA sequence but greatly affect gene expression and impact cellular function). They suggest that while epigenome editing in crop improvement has yielded encouraging results, only the future will tell if it fulfils its great promise in basic research and crop improvement. The CRISPR/Cas9 system has been indicated as the technique of choice for epigenome editing, too.

In Chapter 6 (Bioinformatics and Plant Breeding), Dr Anderson and colleagues provide an extensive list of molecular-marker tools and databases for plant breeders. They discuss genomic selection, marker-assisted selection, real-time genotyping in the field, deep learning (refers to a wide range of statistical methods to identify trends and patterns within large and complex datasets), genome-editing tools (e.g. CRISPR/Cas9) to reduce breeding cycle and pan-genomics.

In Chapter 7 (Bioinformatics Approaches for Pathway Reconstruction in Orphan Crops – A New Paradigm), Drs Evans and Joshi provide an extensive coverage of application of bioinformatics to orphan crops, including sugarcane. The authors have outlined approaches for gene assembly and gene annotation of orphan crops that allow for sequence assembly even when no closely related sequence is available. They have demonstrated the utility of full text mining for gene annotation and pathway discovery. Using *Digitaria exilis* as an example, they have shown that the systems designed for sugarcane can be applied to any orphan crop.

In Chapter 8 (Advances in QTL Mapping and Cloning), Drs Dharminder Bhatia and Darshan S. Brar first provide a glimpse into the early history of quantitative genetics. They cover use of molecular markers for QTL mapping, methods of QTL mapping, cloning of QTL, high-throughput genotyping and phenotyping, genome-wide association mapping, and meta-QTL analysis. This chapter could be helpful for students.

Chapter 9 (Genotype–Environment Interaction and Stability Analyses: An Update), written by me, is an updated version of the article 'Genotype–Environment Interaction: Progress and Prospects' published in the first edition of this book. A distinction between genotype-by-environment interaction and genotype-by-environment correlation has been added. Other additions deal with terms such as 'envirotyping' and climate change. Many new papers, especially those published in the past 5 years, have been reviewed in the updated chapter.

Chapter 10 (Biplot Analysis of Multi-environment Trial Data) represents an updated version of the chapter of Drs Yan and Hunt published in the first edition of this book. They have added much new information from articles published on GGE biplot analysis since 2002. The progress in and use of GGE biplot methodology by applied breeders/agronomists continue apace.

In Chapter 11 (Design and Analysis of Multi-year Field Trials for Annual Crops), Drs Arief, DeLacy and Basford point out that multi-environment trials (METs), being a major component of plant breeding programmes, should use the most appropriate experimental design to maximize their power in predicting genotype performance. They recommend that row-column designs be adopted as standard for METs. They further suggest that multi-year analyses will provide better prediction of genotype performance, estimates of genotype-by-year and genotype-by-year-by-location interactions, and reasonable estimates of variance components.

Chapter 12 (Advances in the Definition of Adaptation Strategies and Yield-stability Targets in Plant Breeding) is an updated version of Dr Paolo Annicchiarico's article titled 'Defining Adaptation Strategies and Yield-stability Targets in Breeding Programmes' published in the first edition of this book. He has reviewed many recently published articles in this chapter.

In Chapter 13 (Prediction with Big Data in the Genomic and High-throughput Phenotyping Era: A Case Study with Wheat Data), Dr Pérez-Rodríguez and colleagues analysed high-dimensional data from CIMMYT's wheat breeding programme, which included more than 45,000 wheat lines that were genotyped using dense single nucleotide polymorphism (SNP) markers and had existing pedigree records. They predicted the performance of unobserved lines using linear models that incorporated markers, pedigree, and the interaction between genotype and environment. Through the new approach, predictions are made on the basis of recommender systems that are routinely used in e-commerce, marketing, biology and, fairly recently, in genomic selection.

In Chapter 14 (Quantitative Genetics in Improving Root and Tuber Crops), Dr Hernán Ceballos has highlighted the role of quantitative genetics in improving important traits, such as fresh root yield (FRY) and dry matter content (DMC), in root and tuber crops. Studies conducted by his group helped them conclude that genomic selection would not be as effective in increasing FRY as it would be for DMC.

In Chapter 15 (Genomic Selection in Rice: Empirical Results and Implications for Breeding), Dr Ahmadi and colleagues dwell on genomic prediction in rice. They address issues related to training population for making selection decisions in the context of pedigree breeding, accounting for genes/QTL involved in the determination of complex traits, as well as for genotype-by-environment interactions. They discuss a strategy for the implementation of genomic selection relative to pedigree breeding.

In Chapter 16 (Novel Breeding Approaches for Developing Climate-resilient Rice), Dr Sandeep Chapagain and colleagues have discussed, at length, the development of new rice varieties that are tolerant to multiple abiotic stresses. They have covered genomics selection, marker-assisted selection, genome-wide association studies, 'omics' and genome-editing approaches to improve rice with respect to biotic and abiotic stresses brought on by climate change. They envision that the use of genome-editing tools for crop improvement will accelerate in the future.

In Chapter 17 (Quantitative Genetics, Molecular Techniques and Agronomic Performance of Provitamin A Maize in Sub-Saharan Africa), Dr Badu-Apraku and colleagues point out that maize consumed in Africa is deficient in nutritional quality, especially lacking in the amino acids lysine and tryptophan, minerals and vitamin A. They highlight the achievements of their maize breeding programme in improving the nutritional quality of maize in sub-Saharan Africa, especially with respect to provitamin A and minerals, using traditional and genomic technologies.

In Chapter 18 (Developments in Genomics Relative to Abiotic Stress-tolerance Breeding in Maize During the Past Decade), Dr Maryke T. Labuschagne discusses the progress in genomics in maize breeding, especially with respect to drought tolerance. She points out that significant research has been done, especially in Africa, where maize is the major staple crop for millions of people. She expects genomics-assisted breeding to speed up the breeding process to develop climate-resilient maize genotypes for production by small-scale farmers and communities.

In Chapter 19 (Exploiting Alien Genetic Variation for Germplasm Enhancement in *Brassica* Oilseeds), Drs Mehak Gupta and S.S. Banga present a very comprehensive treatment of the use of alien variation in improving *Brassica* oilseeds. They present a wealth of information for *Brassica* breeders on wild relatives of *Brassica* as potential sources of desirable traits, introgressed traits through somatic hybridization, sources of cytoplasmic male sterility, and monosomic and disomic alien addition lines in *Brassica*. They discuss use of embryo rescue and protoplast fusion techniques for producing wide hybrids.

In Chapter 20 (Biofortified Pearl Millet Cultivars Offer Potential Solution to Tackle Malnutrition in India), Dr Govindaraj and colleagues point out that one-third of the global population suffers from one or more micronutrient deficiencies (hidden hunger) and that more than 50% of children and women in 20 states of India are anaemic. Biofortification means breeding micronutrient traits into staple food crops, which impact the health of consuming populations. Pearl millet, which is highly

nutritious, is the most important drought- and climate-resilient cereal crop, having high protein, micronutrients and a more balanced amino acid profile than other staple cereals. In 2018, the Government of India renamed millets, including pearl millet, as Nutri-Cereals. The principal emphasis of pearl millet biofortification is on improving primarily grain Fe.

I am grateful to Dr Bruce Walsh for writing the Foreword for this edition, highlighting the importance of new technological tools for plant breeders. I thank all the authors for their cooperation in completing this new volume of *Quantitative Genetics, Genomics and Plant Breeding*.

My thanks go to Dr David Hemming, Acquisitions Editor at CABI Publishing, for his encouragement and support. I also acknowledge other staff members of CABI, especially Ms Emma McCann and Ms Marta Patiño, who efficiently handled the production phase.

I trust the second edition of *Quantitative Genetics, Genomics and Plant Breeding* will be as well received by researchers and teachers as the first edition has been.

2 October 2019

Manjit S. Kang
Adjunct Professor (Quantitative Geneticist)
Department of Plant Pathology
Kansas State University
Manhattan, Kansas, USA

1 Vignettes of the History of Genetics

George P. Rédei
University of Missouri, 3005 Woodbine Ct, Columbia, MO 65203, USA

Mendelism

Although many people consider the beginning of genetics to be the publication of the 'Versuche über Pflanzenhybriden' by Gregor Mendel (Fig.1.1) in 1866 or the submission of the manuscript during the preceding year, the beginning of genetics goes back to thousands of years before.

All geneticists and practically everybody else agree today that Mendel's discovery was an extraordinary achievement. Fewer people know some interesting details about how Mendel achieved it. Not only had he chosen simple characters of an autogamous plant and counted the segregating offspring, but also it was particularly smart that for some he did not have to grow the second generation because the segregation was already evident by inspecting the pods. The circumstances also taught him common sense since he had about 245 m^2 of nursery space in the monastery garden. It also shows that not only was Mendel a very smart man, he also had great sense for practical matters. During his teaching and priestly duties, he also founded a savings and loan bank and a fire brigade. Mendel studied a beautiful *Fuchsia* inflorescence but intuitively did not pursue this ornamental plant further! The chromosome numbers of fuchsias vary a great deal ($2n$ = 22, 55, 66 and 77) and this confused other students of inheritance before and after Mendel.

Mendel himself never claimed any laws to his credit. The term (actually rule (*Regel*) rather than laws) was first used by Carl Correns (1900), and he named them: '1. Uniformitäts- und Reziprozitätsgesetz, 2. Spaltungsgesetz, 3. unabhängige Kombination', namely, first law: uniformity of the F_1 (if the parents are homozygous) and the reciprocal hybrids are identical (in the absence of cytoplasmic differences); second law: independent segregation of the genes in F_2 (in the absence of linkage); and third law: independent assortment of alleles in the gametes of diploids. Thomas Hunt Morgan (1919) also recognized three laws of heredity: (i) free assortment of the alleles in the formation of gametes; (ii) independent segregation of the determinants for different characters; and (iii) linkage–recombination. In some modern textbooks only two Mendelian laws are recognized, but this is against the tradition of genetics in which the first used nomenclature is upheld.

Mendel was a former student and teaching assistant of C.J. Doppler, the physicist, and in the laboratory in Vienna they were already teaching some statistics. Mendel was also fortunate in not finding linkage, which might have been confusing. He used seven characters and obtained 128 (2^7) combinations. Peas have seven linkage groups, thus the probability of independence would have been $6!/7^6 = 720/117{,}649 \approx 0.0061$. Actually some of the genes he studied were syntenic, e.g. *v*, *fa* and *le* in chromosome 4. But the distance between

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

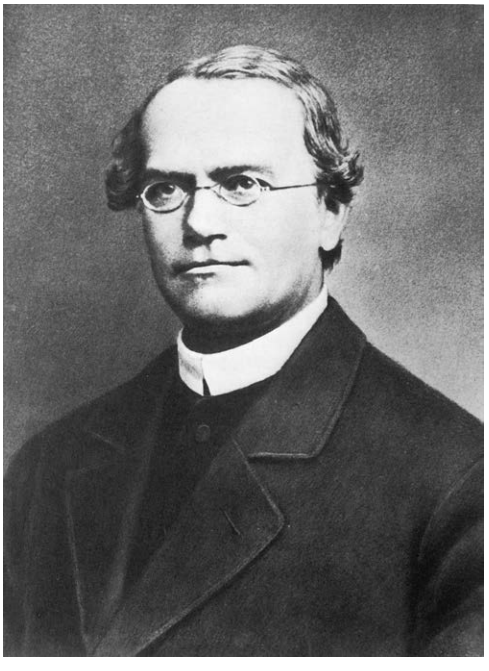


Fig. 1.1. Gregor Mendel, Wikipedia, the free encyclopedia {{PD–US–expired}}.

fa and *le* is 114 map units and *i* and *a* in linkage group 1 (204 map units) are so far away in the chromosome that they segregate independently. It seems that, among the hybrid combinations he had, *v–le* (12 map units) was not included (Blixt, 1975). This was dubbed appropriately 'Mendel's luck', presumably by J.P. Lotsy, a German geneticist of the early 20th century.

The printer, who introduced numerous small errors, had already abused the classic paper of Mendel. The editor took liberties, too, and changed some of the spellings preferred by Mendel. It is known that Mendel corrected by hand at least some of the 40 reprints he received. Only four of these reprints have survived. One of them, sent by Mendel to the renowned Austrian botanist Anton Kerner, was not opened, as revealed by the uncut edges of the paper (Křiženecký and Němec, 1965).

It was quite unfortunate that his contemporaries failed to recognize the significance of his research. Carl Wilhelm Nägeli, the famous professor of botany at the University of Munich and an internationally renowned authority, felt that it was inconceivable that the plants should obey statistical rules. He advised Mendel: 'You should regard the numerical expressions as being only empirical because they cannot be proved rational.' He went even further and suggested to Mendel the study of *Hieracium* apomicts and raised self-doubts in Mendel as to whether the observations he carefully and conscientiously made would really have general validity (Nägeli, 1867).

One should not be entirely negative about Nägeli. He was probably the first who sighted chromosomes around 1842 and described them in German as *Stäbchen* or little sticks in English (Geitler, 1938).

It was not until 1873 that A. Schneider observed mitosis in Platyhelminthes and, 2 years later, Edouard Strasburger reported chromosome numbers for several plant species. Some counts were correct, some not. The term chromosome was coined in 1888 by the surgeon W. Waldeyer, who was not really an experimental biologist but was very good at pigeon-holing (Rédei, 1974).

Professor Nägeli can really be called an expert by the definition of Henry Ford, who said the expert knows what cannot be done: even when he sees that it has already been accomplished, he can also explain why it should not have been successful. Nägeli almost shot down the Mendelian results. He might also have been influential on Wilhelm Olbers Focke, who in 1881 in his monograph on plant hybrids refers only 15 times to Mendel (nine times in connection with *Hieracium* but only once about the pea experiments) but mentions the name and work of Gärtner 409 times, Kölreuter 214 times and several others dozens of times.

Nägeli evoked the ire of the medical researchers with his ideas on bacterial pleomorphism. Pleomorphy meant that bacteria (he called them *Schizomycetes*) were not supposed to possess hard heredity. He believed that their variability is not hereditary but depends entirely on the culture conditions. Apparently, his laboratory skills were insufficient and he did not understand what pure cultures are. Unfortunately, his influence and 'authority' were a serious impediment to the development of bacteriology.

Dr W. Migula, professor at the College of Technology in Karlsruhe, Germany, gives a vivid account about the situation in his *System der Bakterien* in 1897:

> When Nägeli says, p. 20, that 'Cohn [the founder of modern bacterial systematics in 1872] had established a system of genera and species, in which each function of the *Schizomycetes* [bacteria] is represented by a particular species; by this he expressed the rather widespread view exclusive to physicians. So far I have not come across any factual ground that could be supported by morphological variations or by pertinent definitive experiments.' When Nägeli still says this in 1877, one must either assume that he was unaware of the work of the preceding 5 years, or that he chose to ignore it on purpose because it did not fit his theory.

Nägeli has also some positive legacies. I have mentioned before that he was probably the first to report seeing chromosomes. In 1884, he published a large volume entitled: *Mechanisch-physiologische Theorie der Abstammungslehre*, which is also the first systematic effort to create a molecular interpretation of the hereditary material.

Mendel's problems did not cease with his death. Anselm Rambousek, who succeeded Mendel as abbot of the monastery, destroyed a large part of the unpublished records and personal notes after the death of his predecessor. There are different ways of leaving a historical legacy.

Fortunately, Mendel did not live to read Sir Ronald Fisher's (1936) devastating criticism. Fisher, one of the greatest statisticians ever lived, questioned, in good faith, the 'too good to be true' data of Mendel – although Fisher tried to find excuses for Mendel, such as an assistant who was familiar with his expectations and might have deceived him, or that he figured out what he was supposed to find and just wanted to demonstrate the validity of his hypothesis. Nobody will ever find out what happened. Some of the sensation-hungry public media periodically revisit the Fisher paper and question Mendel's integrity. His principles are beyond doubt. I do not wish to go into the details because these are familiar to the majority of the students and workers in genetics. Alfred Sturtevant (1965) points out that Fisher erred in the dates, in the number of years of the experiments and misrepresented some of the statements in Nägeli's letters to Mendel.

F. Weiling (1966), a German statistician, after a thorough analysis arrived at similar conclusions. Weiling also used more technical arguments. He pointed out that the pollen tetrads may clump and then the distribution may be biased and suggests the following calculations for chi-square:

$$\chi^2 = \frac{(x - Np)^2}{Np(1-p)}$$

where x = the observed, say, recessives, N = the number of individuals in the sample, p = the expected frequency of the phenotype. Weiling provides the following hypothetical example: $x = 152$, $N = 580$, $p = 0.25$:

$$\chi^2 = \frac{(152 - [580 \times 0.25])^2}{580(0.25) \times (0.75)} = \frac{(152-145)^2}{108.75}$$
$$= \frac{49}{108.75} = 0.4505747126$$

This has a probability that is very different from that calculated by Fisher. Weiling also claims that Fisher erred by assuming the identity of the reciprocal crosses and did not take it into account and that might have affected the chi-square, which should have been calculated using a correction factor c:

$$\chi^2 = \frac{(x - Np)^2}{c(Np[1-p])}$$

If the distribution is not really binomial but semi-random, this also affects the chi square value.

Not being a statistician, I do not want to take a position in the dispute. I only wish to provide some food for thought in this case or in general. One point is indisputable: no matter how Mendel reached his conclusions, he was right. Back in the 1950s, I conducted larger experiments with monogenic segregation of auxotrophic mutants of *Arabidopsis* and observed an even better fit to the 3 : 1 under axenic conditions.

Psychologists have a term for problems of judgement: multistability of perception. In layman's words, you see what you want to see. Of course, you do not always get what you see. The British artist Gerald H. Fisher (1968) (I do not know whether he was kin to Sir Ronald) graphically illustrated how these things happen (Fig. 1.2). The upper drawing shows an ugly man, the lower figure displays an undressed woman but if you look long enough both pictures show the same.

Sometimes, failing memory or perhaps a drive for humour distorts the historical facts. In 1949, R.C. Punnett reminisced on the origin of the Hardy–Weinberg law and said:

Fig. 1.2. Gerald Fisher's (1968) graphic illustration of the multistability of perception. Basically the same object (or principle in science) may mean different things depending on when and how one looks at it. Both figures above may appear as a sad male face or a nude. (By permission of the Psychonomic Society.)

> I was asked why it was that, if brown eyes were dominant to blue, the population was not becoming increasingly brown eyed: yet there was no reason for supposing such to be the case. I could only answer that the heterozygous browns also contributed their quota of blues and that somehow this leads to equilibrium. On my return to Cambridge I at once sought out G.H. Hardy with whom I was then very friendly. For we had acted as joint secretaries to the Committee for the retention of Greek in the Previous Examination and we used to play cricket together. Knowing that Hardy had not the slightest interest in genetics I put my problem to him as a mathematical one. He replied that it was quite simple and soon handed to me the now well-known formula pr = q2 (where p, 2q and r the proportions of *AA*, *Aa* and *aa* individuals in the population varying for the *A*–*a* difference). Naturally pleased at getting so neat and prompt an answer I promised him that it should be known as 'Hardy's Law' a promise fulfilled in the next edition of my *Mendelism*. Certain it is that 'Hardy's Law' owed its genesis to a mutual interest in cricket.

Punnett might not have ever read the seminal paper of Hugo de Vries in 1900, where he said much earlier:

> Si l'on appelle D les grains de pollen ou les ovules ayant un caractère dominant et R ceux qui ont le caractère récessif, on peut se représenter le nombre et la nature des hybrides par la formule représentative suivante, dans laquelle les nombres D et R sont égaux:
>
> $$(D+R)(D+R) = D^2 + 2DR + R^2$$

This is, of course, no different from what all textbooks call either the Hardy–Weinberg law or the Castle–Hardy–Weinberg theorem.

Why Genetics was a Late Bloomer

The question often emerges why genetics started so late relative to other sciences. Copernicus (1473–1543) centuries earlier had proposed essentially valid ideas about the celestial bodies. Galileo (1564–1642) developed theories on dynamics and astronomy. Newton (1642–1729), who understood something about genetics by being also a sufferer from the complex hereditary disease gout, pioneered in gravitation and energy. Dalton (1766–1844) developed an atomic theory,

although he was afflicted by X-linked red–green colour blindness and, being a physicist, he quite clearly described his malady. In literature, Shakespeare (1564–1616), Molière (1622–1673) and Goethe (1749–1832) preceded Mendel. The latter – besides being an immortal poet – contributed significantly to the understanding of the biology of development. Mozart (1756–1791) and Beethoven (1770–1827) elevated music to an unsurpassable beauty. Strangely, Beethoven was tormented by a hearing deficit and that might have been the reason why he elected not to marry and have offspring, although he was romantically involved with several women.

There were several causes of the late development of genetics. Basic biological mechanisms of reproduction were not understood. Experimental procedures were not used. I cannot tell whether the ancient Egyptians comprehended the consequences of human inbreeding, but the artists of the 14th century BC depict the pharaoh and his wife's offspring like Wilhelm Johannsen's (1857–1927) famous beans (Fig. 1.3).

Aristotle (384–322 BC) writes that, in Abyssinia, mice get pregnant if they lick salt. He probably did not believe it, but the 'information' might have come from a respected source so he felt obligated not to dispute it. He also stated that women had fewer teeth than men. It is hard to understand why he never looked into the mouth of his wife or mother; this would not have required a grant or special equipment.

When Aristotle reviews the ancient theories of sex determination, he finds them all unsatisfactory:

> Some suppose that the difference [between sexes] exists in the germs from the beginning; for example, Anaxagoras and other naturalists say that the sperm comes from the male and that the female provides the place [for the embryo], and that the male comes from the right, the female from the left, since in the uterus the males are at the right and the females at the left. According to others, like Empedocles, the differentiation takes place in the mother, because, according to them, the germs penetrating a warm uterus become male, and a cold uterus female.

Several of his other reported cases of heredity seem, however, quite plausible and sensible, while others are utter nonsense. Aristotle states that mutilations are not transmitted to the offspring but blindness and some scars may be. There are more defective males than females. The normal eye colour is black, and blue is a deficiency of the shade. Some of the travelling 'Freiherr von Münchhausen'-like stories find their

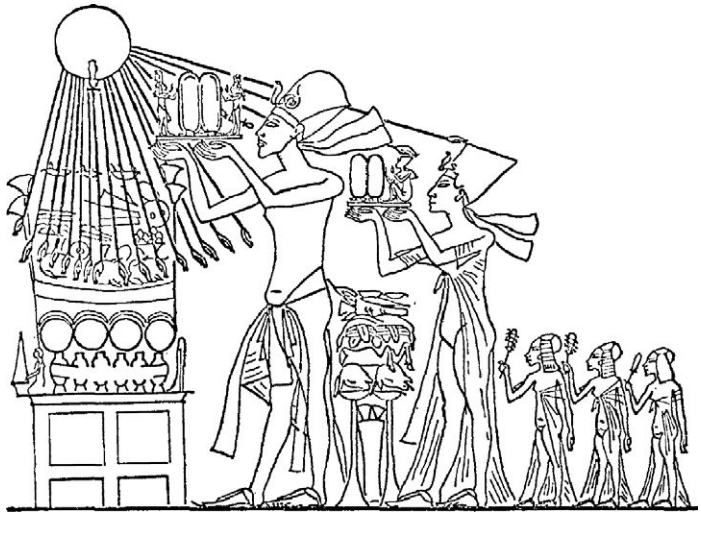


Fig. 1.3. King Akhenatan, Queen Nefertiti and three of their daughters. Bas-relief from the tomb of Apy at Amarna, c. 1362 BC. (Drawn by Cyril Aldred after Davies. Aldred, C. (1961) *The Egyptians, Ancient People and Places*, Thames and Hudson, New York. By permission.)

ways into his erudite books. In Libya – he writes – because of drought and heat diverse species of thirsty animals congregate at an oasis and mate. From such misalliances, for example, camel × sparrow → ostrich arises or the wild boar would have its origin by ants mating with lions. The Roman Pliny (AD 23–79) remarks 'si libeat credere' – if we are permitted to believe in tall stories.

On the other hand, even students of Linnaeus – for example, the savant Austro-Finlandus Johannes J:nis Haartman (1751) – faithfully retell the incredible fantasies. So does practically everybody else through the centuries. The Soviet charlatans during the Lysenko era in the 20th century (Medvedev, 1969), who destroyed genetics and maimed many outstanding geneticists (e.g. Agol, Vavilov and hundreds of others), postulated similar fantastic nonsense (vegetative hybrids, inheritance of environmentally acquired traits, etc.).

Besides the lack of experimentation and the slavish submission to the ancient books, there was another negative force, expressed by Joshua Sylvester in the 16th century. Sylvester answers the 'New objection of Atheists, concerning the capacitie of the Ark':

> O profane mockers! if I but exclude
> Out of this Vessell a vast multitude
> Of since-born mongrels, that derive their birth
> From monstrous medly of Venerian mirth:
> Fantastick Mules, and spotted Leopards
> Of Incest-heat ingendred afterwards:
> So many sorts of Dogs, of Cocks, and Doves
> Since, dayly sprung from strange & mingled loves,
> Where in from time to time in various sort,
> Daedalian Nature seems her to disport:
> If plainer, yet I prove you space by space
> And foot by foot, that all this ample place,
> By subtill judgement made and Symmetrie,
> Might lodge so many creatures handsomely,
> Sith every brace was *Geometricall*:
> Nought resteth (*Momes*) for your reply at all;
> If, who dispute with God, may be content
> To take for current, Reason's argument.

The Reverend Dr Hodge of Princeton University, expressing the opinions of many of his contemporaries about Darwinism, remarked: 'to ignore design as manifested in God's creation is to dethrone God' (Provine, 1971, p. 10).

Dr A.F. Wiegmann, a physician from Brauschweig, Germany, was a prize-winner of the Physical Section of the Royal Prussian Academy of Sciences in 1826; his thesis in the competition sought to shed light on the problem: 'Gibt es ein Bastarderzeugung im Pflanzenreiche?' (Is there any hybridization in the plant kingdom?). On the second attempt he received only half the prize because he could not prove to the distinguished panel's complete satisfaction that plants do form hybrids. In his detailed report, he complains about his deteriorating vision, trembling hand, difficulties in bending and kneeling in his backyard, and, above all, he is worried about the neighbours who might think that he is sodomizing plants (Roberts, 1965).

Some of the attempts with animal hybridization (wolf × mastiff) described by George Louis Le Clerc Compte de Buffon (1707–1788) were even more disastrous. The wolf killed the dog and mauled the curious experimenter (Olby, 1966).

During the preceding era, experimentation had not been very popular. All this was changing now with the Enlightenment philosophy of the 18th century. The language may still be Latin but the ideas are revolutionary. In 1759 the St Petersburg Russian Academy of Sciences offered a prize for proving:

> Sexu plantarum argumentis et experimentis novis, praeter adhuc iam cognita, vel corroborare vel impugnare, premissa expositione historica et physica omnium plantae partium, qui aliquid ad fecundationem et perfectionem seminis et fructus conferre creduntur. [Sexuality of plants should be confirmed or refuted by arguments and new experiments, besides those that are already known, by presenting the history and the physical parts of all plants that are believed to have contributed to the seed and fruits.]

Kölreuter, an early plant hybridizer, apparently, stipulated these requirements (Roberts, 1965).

This is a major milestone on the way to experimental science. The Academy wanted to see not just the records of the observations but also the physical evidence, fruits, seeds and all other plant parts. Linnaeus entered and won the contest and later expressed his wishes to spend the rest of his life studying plant hybrids.

Felix Hoppe-Seyler, a not particularly modest editor of the journal *Hoppe-Seylers Medizinische-Chemischen Untersuchungen*, set similar critical requirements. When he received Friedrich Miescher's manuscript of the initial study on nuclein in 1869, the thorough editor did not

publish it until 1871, when he himself had a chance to confirm the information along with two separate papers, authored by two of his students, which showed that Miescher was correct. Actually Hoppe-Seyler and his team had proved that nuclein was not a substance unique to pus cells but was present in red blood cells, in yeast and even in milk, and this is also the beginning of the DNA story (Borek, 1965).

The obvious question arises: is such an editorial policy desirable or not? In this case it actually worked well and eventually the priority was posthumously credited to Miescher alone, despite the 'piracy' of his intellectual property. Editorial heavy-handedness does not always have such a happy ending. Hoppe-Seyler rejected the paper of MacMunn dealing with haematin, a pigment present in tissues besides blood. MacMunn's results were thus not appreciated until 1925, when another biochemist, Keilin, showed that MacMunn was right and this pigment was important for respiration (Borek, 1965).

There are several examples of similar poor judgement by experts. The editor of the *Lancet* rejected – for lack of understanding – the seminal manuscript of L. and H. Hirszfeld on the frequencies of the three alleles of the ABO blood group and the article could find its way only into *Anthropologie*, a less widely read journal (Stoneking, 2001). H.J. Muller was fired from the University of Massachusetts shortly before he was awarded the Nobel prize (1946) because the administration was not satisfied with his teaching skills. A graduate student, according to my non-scientific survey, had a completely different view. *Nature* (*London*) rejected the manuscript of Hans A. Krebs, who became a Nobel laureate for the same work in 1953. In 1970, a distinguished genetics panel declared *Arabidopsis* to be *planta non grata*, but, by 2000, it became the first completely sequenced higher plant and more papers are being published about it than any other plant species.

Hugo de Vries, Carl Correns and Erich von Tschermak-Seysenegg rediscovered Mendel's work in 1900. The circumstances of the rediscovery were also controversial. H. de Vries, in this first paper, did not refer to Mendel and his explanations regarding whether he had ignored or forgotten him are contradictory. In a letter written to H.F. Roberts (1965), de Vries claimed that he worked out the Mendelian rules all by himself without the help of Mendel's work. A.H. Sturtevant (1965, p. 27) casts some doubt on the truthfulness of this claim:

> In 1954, nineteen years after the death of de Vries, his student and successor Stomps reported that de Vries had told him that he learned of Mendel's work through receiving a reprint of the 1866 paper from Beijerinck, with a letter saying that he might be interested in it. The reprint is still in the Amsterdam laboratory, as has been stated.

Despite these facts, de Vries generally receives more credit in the literature than Correns, whose contributions to genetics are much more substantial. Tschermak's work is the least valuable and the least original.

Bateson, while travelling on a train and reading, came across the Mendelian experiments and the confirmations. He became the most ardent Mendelian and the most diligent public relations man for the new ideas. He encountered stiff resistance from various corners, mainly from the biometricians, students and followers of Sir Francis Galton. Bateson published an enthusiastic book in 1902: *Mendel's Principles of Heredity: a Defence*.

Karl Pearson, a man of enormous intellect, was one of the most vociferous critics of Bateson. According to him, the purity of the gametes theory was 'not elastic enough to account for the numerical values of the constants of heredity hitherto observed' (Pearson, 1904). He requested that the Mendelians provide 'a few general principles . . . which embrace all the facts deducible from the hybridization experiments' (Pearson, 1904). Bateson was ill equipped to deal with the mathematical tasks that would 'form the basis of a new mathematical investigation' (Pearson, 1904). G. Udney Yule (1907) came to the rescue of Batesonism by accepting the compatibility of Mendelism and biometry. Wilhelm Johannsen (1909) wrote a great book with the purpose of demonstrating the need for biometry in understanding genetics. It is regrettable that this monumental work has not been translated into English and is inaccessible to many geneticists due to a language barrier.

The amalgamation of biometry and genetics did not happen readily. In the journal *Genetics*, the statistical papers are still relegated to the back of issues. Many geneticists find the language and

concepts obtrusive because of lack of adequate mathematical preparation. Roger Milkman reported several years ago about an international meeting of statistical genetics, that the papers were apparently beautiful, albeit he did not understand them but hoped that the speakers did.

Pearson's confidence in the application of biometry to genetics was well vindicated by the development of the shotgun sequencing of genomes, which could not have been carried out without very powerful computers and computer programs (Sharing the glory not the credit. *Science* 291, 1189 (2001)).

The general acceptance of Mendelism continued after the rediscovery not only by the biometricians but also by the embryologists, evolutionists and zoologists. Nevertheless, at the 6–8 January 1909 meeting of the American Breeders' Association in Columbia, Missouri, Professor T.H. Morgan of Columbia University did not attend personally – maybe because of contempt for the predominantly agricultural audience – but he submitted a paper entitled 'What are "factors" in Mendelian explanations?' A member of the Zoology Department read it:

> In modern interpretation of Mendelism, facts are being transformed into factors at a rapid rate. If one factor will not explain the facts, then two are invoked; if two prove insufficient, three will sometimes work out. The superior jugglery sometimes necessary to account for the results may blind us, if taken too naïvely, to the common-place that the results are often so excellently 'explained' because the explanation was invented to explain them. We work backwards from the facts to the factors, and then, presto! explain the facts by the very factors that we invented to account for them. I am not unappreciative of the distinct advantages that this method has in handling the facts. I realize how valuable it has been to us to be able to marshal our results under a few simple assumptions, yet I cannot but fear that we are rapidly developing a sort of Mendelian ritual by which to explain the extraordinary facts of alternative inheritance.

The Rise of *Drosophila* and Cytogenetics

By the time this paper and others similar in tone appeared in print, an unusual, strange event took place. (I am relating the story as I heard it from Dr E.G. Anderson, who was at that time a graduate student of R.A. Emerson at Cornell University.)

C.W. Woodworth, an entomology student, introduced *Drosophila* to the Harvard laboratory of William Castle, and Morgan also used it as a tool for his embryology class. One day, he wanted to demonstrate the phototropism of the flies. As Mrs Lillian Morgan opened a matchbox containing *Drosophila,* Professor Morgan went to the window and told the students to watch how the flies would come towards him. Facing the flies, Dr Morgan discovered a white-eyed one. He became interested in it, but, despite the assistance of the students, the fly escaped. Next day, a mutant male was captured and thus the future of genetics was changed.

In 1910 and 1911, Morgan, an embryologist, published the first genetics paper on 'sex-limited' inheritance. This was new for *Drosophila* and Morgan but not for genetics. Four years earlier, Doncaster and Raynor (1906), working with the *Abraxas* moth, discovered criss-cross inheritance and, despite the assistance of William Bateson, the puzzle could not be rationalized. Their hypotheses broke down.

Miss N.M. Stevens and Professor Edmund Wilson each showed in 1905 that the 'unknown' X chromosome of Henking (1891) was actually a sex-determining chromosome. Wilson and Morgan were colleagues at Columbia University and they knew about each other's work. Thus, sex linkage was a simple inference.

There was, by that time, a lot of interest in chromosomes. Before the turn of the century, several authors had published chromosome numbers, including that of humans. Bardeleben observed about 16, while Flemming was sure that there were more than 16 (Sutton, 1903). De Winiwarter (1912) in sectioned testes observed 46 autosomes + an X chromosome but no Y chromosome. The latter is, of course, the smallest: according to the human genome draft (excluding gaps) it contains only 21.8 megabases versus the X chromosome, which has 127.7 megabases (Lander *et al.*, 2001). In the ovaries, de Winiwarter observed, correctly, a total of 46 chromosomes. During the following decades, various numbers were reported even by the same investigators (von Nachtsheim, 1959). In 1952, T.C. Hsu (von Nachtsheim, 1959), using a hypotonic solution, claimed 48, but subsequently Tijo and Levan (1956) showed, by a

similar technique, adding also colchicine, beyond any doubt that humans have only 46 (von Nachtsheim, 1959).

A historical irony is that, in 1953, Cyrill Darlington, one of the most renowned cytologists, published a popular book *Facts of Life* with a photomicrograph of Hsu on the cover and showing only 46 chromosomes, but he cited it as evidence for 48 human chromosomes (von Nachtsheim, 1959).

The problem remained controversial, although the majority of cytologists confirmed that 46 was the correct number. M. Kodani in several papers between 1956 and 1958 reported 46, 47 and 48 chromosomes in both Japanese and US white individuals (von Nachtsheim, 1959).

Various banding techniques were developed during the 1970s by Torbjörn Casperson and associates and were expanded by others, which yielded the human chromosome pictures as they are used for cytogenetic maps (Caspersson *et al*., 1968). By 1996, Speicher *et al*., using multiplex fluorescence *in situ* hybridization (FISH) technology, distinguished each human chromosome with a distinct colour.

Let us jump back in time to 1903, when Walter Sutton published an epoch-making paper on chromosomes in heredity. He correctly asserted that the chromosomes are not separated by paternal and maternal groups, although the two groups are equivalent. There are two distinct types of nuclear divisions, equational and reductional (van Beneden, 1883). The chromosomes retain their individuality in the process. He assumed with Bardeleben that there are 16 chromosomes in humans and thus they may produce $16 \times 16 = 256$ gametic combinations. The 256 gametic types can thus produce $256 \times 256 = 65{,}536$ phenotypes. He assumed linkage, but for recombination he suggested 'segmental dominance'. His combinations are not too far from the current estimated human gene numbers.

Carl Correns, who also discovered cytoplasmic (chloroplast) inheritance, observed linkage in 1900 and suggested in 1902 a model for recombination 9 years before Morgan.

The majority of geneticists know that Carl Correns was one of the three rediscoverers of the Mendelian principles in 1900 and reported linkage in *Matthiola* in 1900. He was also one of the discoverers of cytoplasmic inheritance (Correns, 1909).

In 1902, Correns suggested a mechanism for crossing over 9 years before Morgan's paper appeared in the *Journal of Experimental Zoology*.

> We assume that in the same chromosome the two anlagen of each pair of traits lie next to each other (A next to a and B next to b, etc.) and that the pairs of anlagen themselves are behind each other. A, B, C, D, E, etc. are the anlagen of parent I; a, b, c, d, e, etc. are those of parent II. Through the usual cell and nuclear divisions the same type of products are obtained as the chromosomes split longitudinally… When one pair contains antagonistic anlagen, while the rest of the pairs are formed of two identical types of anlagen, or the anlagen are 'conjugated' as they are in *Matthiola* hybrids, which I have described, then further assumptions are necessary… Then AbCdE/aBcDe and aBcDe/AbCdE yield both AbCdE and aBcDe; ABcdE/abCDe and abCDe/ABcdE both ABcdE and abCDe, etc.

Another really remarkable paper is slowly sinking into oblivion or is totally misrepresented. On 9 July 1909 (more than two decades earlier than the *Neurospora* work of Carl Lindegren in 1932), F.A. Janssens, professor at the University of Louvain, Belgium, presented his theory of chiasmatypy in the journal *La Cellule*:

> In the spermatocytes II, we have in the nuclei chromosomes, which show one segment of two clearly parallel filaments, whereas the two distal parts diverge… The first division is therefore reductional for segment A and a and it is equational for segment B and b… The 4 spermatids contain chromosomes 1st AB, 2nd Ab, 3rd ab, and 4th aB. The four gametes of a tetrad will thus be different… The reason behind the two divisions of maturation is thus explained … The field is opened up for a much wider application of cytology to the theory of Mendel.

Elof Carlson (1966) – in his otherwise excellent book – cites this paper and even shows with some drawings that Janssens believed that recombination takes place at the two-strand stage. The drawings of Carlson are, however, nowhere in the publication of Janssens. When Morgan discovered crossing over 2 years later, he acknowledged the priority of Jannsens, who, however, had only cytological evidence.

Morgan's student, Sturtevant, constructed the first genetic map and recognized inversions as crossing-over inhibitors. Morgan, Bridges and Muller revealed the basic mechanics of recombination. Bridges discovered non-disjunction, deletion, duplication and translocation. The list above includes only the most significant discoveries of the chromosomal theory of inheritance.

Bateson, the great champion of genetics, who coined the term genetics and whom, in 1926, T.H. Morgan eulogized with these words: 'His rectitude was beyond all praise and recognized by friend and foe alike,' concluded a memorial lecture in 1922 at the University of Pennsylvania with the following warning:

> I think we shall do genetical science no disservice if we postpone acceptance of the chromosome theory in its many extensions and implications. Let us distinguish fact from hypothesis. It has been proved that, especially in animals, certain transferable characters have a direct association with particular chromosomes. Though made in a restricted field this is a very extraordinary and most encouraging advance. Nevertheless the hope that it may be safely extended into a comprehensive theory of heredity seems to me ill-founded, and I can scarcely suppose that on wide survey of genetical facts, especially those so commonly witnessed among plants, such an expectation would be entertained. For phenomena to which the simple chromosome theory is inapplicable, save by the invocation of a train of subordinate hypotheses, have been there met with continually, as even our brief experience of some fifteen years has abundantly demonstrated.
>
> (Bateson, 1926)

Morgan very successfully exploited the potentials of his 'fly room' and trained a remarkable series of students (Bridges, Sturtevant, Muller,* Dobzhansky, Curt Stern, Bonnier, Komai, Gabritchevsky, Olbrycht, Altenburg, Weinstein, Gowen, Lancefield, Mohr, Nachtsheim, E.G. Anderson, Jack Schultz and others), whose work became the foundation of classical genetics and the main menu of textbooks for decades to come. Morgan's association with the California Institute of Technology signalled a more modern trend of genetics and the development of a younger generation of geneticists, such as Beadle,* Tatum,* Ephrussi, Delbrück,* Norman Horowitz, Lindegren, Schrader and E.B. Lewis.* The students of their students, such as Lederberg,* Doerman, Srb and others, made a lasting impact on the future course of genetics.

An interesting episode of the Cal Tech and the preceding period of Morgan has been recorded by Henry Borsook (1956). In the late 1920s, Edwin Cohn, the physical chemist, asked T.H. Morgan, the first Nobel-laureate geneticist, what his research plans were. Morgan's answer was: 'I am not doing any genetics, I am bored with genetics. But I am going out to Cal Tech where I hope it will be possible to bring physics and chemistry to bear on biology.'

Shortly after Morgan arrived at Cal Tech, Albert Einstein visited the laboratory and posed almost the same question. Morgan's answer was about the same as before. Einstein shook his head and said, 'No, this trick won't work. The same trick does not work twice. How on earth are you ever going to explain in terms of chemistry and physics so important a biological phenomenon as first love?' Sure enough, in the 1930s, Morgan could not provide an answer to Einstein's question, but at the current rate of advances of molecular neurogenetics some clues may soon be available.

Mutation

In 1927, H.J. Muller in *Drosophila* and independently L.J. Stadler (1928) in barley and maize proved that X-rays can induce mutations.

The Nobel-laureate immunologist, Peter Medawar, remarked once that wise people may have expectations, but only fools make predictions. Of course, brilliant people may make brilliant errors.

In a somewhat ill-conceived manner, in 1941, at the 9th *Cold Spring Harbor Symposium on Quantitative Biology* (p. 163), H.J. Muller stated:

> We are not presenting ... negative results as an argument that mutations cannot be induced by chemical treatment... It is not expected that chemicals drastically affecting the mutation process while leaving the cell viable will readily be found by our rather hit-and-miss methods. But the search for such agents, as well as the study of the milder, 'physiological' influences that may affect the mutation process, must continue, in the expectation that it still has great possibilities before it for the furtherance both of our understanding and our control over the events within the gene.

Charlotte Auerbach and J.M. Robson might have already solved the problem when belatedly – because of wartime security restrictions – in 1944 they reported successful induction of mutations with radiomimetic chemicals. Muller worked for a period of time along with Auerbach

* Nobel laureates.

in G. Pontecorvo's laboratory in Edinburgh after his return, via Spain, from his unhappy sojourn in the Soviet Union.

Despite all, H.J. Muller was the well-deserving second geneticist recipient of the Nobel prize for his studies on mutation. *Science* magazine in November 1946 (Vol. 104, p. 483) proudly reported the award, and perhaps appropriately with a printing 'mutation' or typo.

Non-nuclear Inheritance

W. Haacke assumed in 1893 that the waltzing–walking traits of mice are located in cytoplasmic elements (the centrosome), whereas coat colour (white–grey) segregation is assured by the reductional division of the chromosomes. 'I do not know whether the number of chromosomes present in mice had been recorded, but this number would enable us to establish the possible combinations.' The fact that he was able to obtain experimentally all 16 combinations of these four traits seemed to indicate to him the validity of this interpretation.

C. Correns (1909) and E. Baur (1909), independently, reported genuine cytoplasmic inheritance in various plants, and their findings were abundantly confirmed later.

Professor T.H. Morgan in 1926 expressed the following view: 'except for the rare cases of plastid inheritance, the inheritance of all known characters can be sufficiently accounted for by the presence of genes in the chromosomes. In a word the cytoplasm may be ignored genetically.'

John R. Preer, Jr (1963), an eminent contributor to the field, remarked:

> Cytoplasmic inheritance is a little bit like politics and religion from several aspects. First of all, you have to have faith in it. Second, one is called upon occasionally to give his opinion of cytoplasmic inheritance and to tell how he feels about the subject.

Pleiotropy

The term pleiotropy was coined by Ludwig Plate, a German geneticist, and he wrote in 1913:

> ein Gen in manchen Fällen gleichzeitig mehrere Markmale, die zu ganz verschiedenen Organen gehören können, beinflußt. Eine solche Erbeinheit habe ich ... pleiotrop genannt [a gene in many instances can influence several traits, which can be involved with different organs].

Interestingly, in Sutton (1959) the following discussion has been recorded:

> *Fremont-Smith:* Can one gene operate only in one highly specified environment and perform only one function? Would any other environment either suppress its activity or be lethal? Or can a gene perform a variety of functions, depending upon the environment to which it is exposed?
> *Lederberg:* There is no qualitative difference in the product, depending on the environment.
> *Wagner:* But that which the gene forms acts differently in different environments.
> *Fremont-Smith:* It has no multiple potentiality at all?
> *Lederberg*: Pleiotropism non est.
> *Fremont-Smith*: Did you add, at the 'dogma' level?
> *Lederberg*: In terms of the primary product, that is the doctrine.

By the 1980s and 1990s, mitochondrial functions had been thoroughly studied by many geneticists. The fact that single base pair mutations in the human mitochondrial $tRNA^{Leu}$ and other tRNAs may cause more than single human disease is clear evidence for pleiotropy (Fig. 1.4).

Definition of the Gene

These and other recent developments may modify the definition of the gene:

> Woltereck (1909): A reaction norm.
> Sturtevant (1965): Mendel usually used the term Merkmal for what we now term gene.
> Suzuki *et al*. (1976): The fundamental physical unit of heredity.
> Klug and Cummings (1983): A DNA sequence coding a single polypeptide.
> Elseth and Baumgardner (1984): A segment of the DNA that codes for one particular product.
> Strickberger (1985): In modern terms, an inherited factor that determines a biological characteristic of an organism is called a gene.
> Russel (1992): The determinant of a characteristic of an organism.
> Gray Lab Internet Glossary (2001): Genes are formed from DNA, carried on the chromosomes and are responsible for the inherited characteristics that distinguish one individual from another. Each human individual has an estimated 100,000 separate genes.

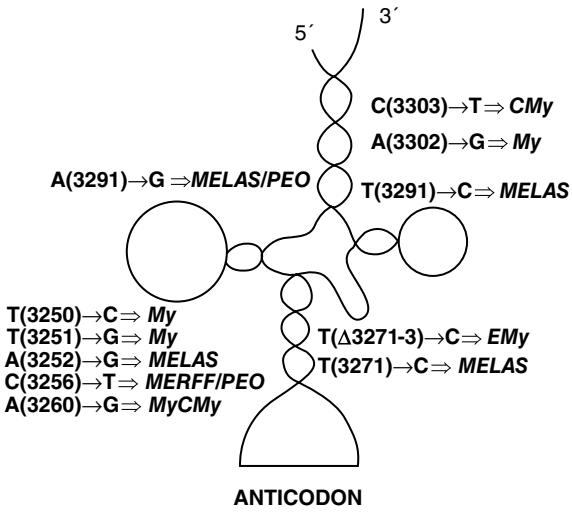


Fig. 1.4. Pleiotropic mutations in a human mitochondrial gene transcribed into UUR-tRNA^{Leu}. *CMy*, cardiomyopathy; *My*, myopathy; *MELAS*, mitochondrial myopathy, encephalopathy, lactic acidosis, stroke; *PEO*, progressive external ophthalmoplegia; *EMy*, encephalopathy, myocardia; *MyCMy*, myopathy, cardiomyopathy; *MERFF*, myoclonic epilepsy, ragged-red fibres. (Redrawn and modified after Moraes, 1998.)

Each of these definitions has some correct elements. Probably the best is still that of Wolterneck (1909). The least pleasant one is the last. Genes are not formed from DNA. Genes are either DNA or RNA depending on the organism. Genes are not on the chromosomes but the genes are in the chromosomes; actually the DNA forms the backbone of the chromosomes. The number of human genes is most unlikely to be 100,000; the latest estimates indicate about 35,000.

How would I define briefly the gene today?

> Gene: a specific functional unit of DNA (or RNA) potentially transcribed into RNA or coding for protein(s). A group of cotranscribed exons but, due to alternative splicing, exon shuffling, overlapping or using more than one promoter or termination signal, the same DNA sequence may encode more than a single protein.
> A common structural organization of protein-encoding genes in eukaryotes:
>
> enhancer – promoter – leader – exons – introns – termination signal – polyadenylation signal – downstream regulators

The vast majority of the human genes are 'mosaics' containing seven to nine exons of 120 to 150 bp each. In some genes, the exon number may be much larger (e.g. in titin about 200). In between exons, there are 1000–3500 bp introns. The size of the introns may be many times larger. The number of coding nucleotides generally varies between 1100 and 1300 bp, but the larger genes may have much longer coding sequences. The exons + introns + 5′ and 3′ untranslated sequences combined, the genomic genes, in general, extend to 14–27 kb DNA. The human dystrophin gene in the X chromosome extends to about 2300 kb. A large fraction of the human genes are alternatively spliced and thus the same gene may be translated into two, three or more kinds of proteins. Genic sequences (2–3%) are richer in GC nucleotides than the non-coding tracts. In prokaryotes, introns are rare and the genes are much smaller (Rédei, 2002).

Gene Numbers

The number of genes per genome of an organism can be estimated by molecular analysis on the basis of mRNA complexity or by total sequencing of the genome. The estimates based on mRNA can be best determined when the entire genome is sequenced. By the latter method, the single-stranded RNA phage, MS2, was found to have four genes. The gene number has also been estimated from mutation frequencies. If the overall induced mutation rate, for example, is 0.5 and the mean mutation rate at selected loci is 1×10^{-5}, then the number of genes is $0.5/(1 \times 10^{-5}) = 50{,}000$. Although this method is loaded with some errors, the estimates so obtained appear reasonable. On the basis of mutation frequency in *Arabidopsis*, the total number of genes was estimated to be about 28,000 (Rédei *et al.*, 1984). The number of genes of *Arabidopsis* was

estimated to be 25,498 after sequencing the genome. In *Drosophila*, ~17,000 genes were claimed on the basis of mRNA complexity. On the basis of the sequenced genome, the estimate is now ~13,600. During the 1930s, C.B. Bridges counted ~5000 bands in the *Drosophila* salivary chromosomes and for many years it was assumed that each band represented a gene. By 1928, John Belling had counted 2193 chromomeres in the pachytene chromosomes of *Lilium pardalinum* and assumed that this number corresponded to the number of genes (Belling, 1928).

Nucleotide sequencing of 69 salivary bands in the long arm of chromosome 2 of *Drosophila* pointed to the presence of 218 protein-coding genes, 11 tRNAs and 17 transposable element sequences within that ~2.9 Mb region. The shotgun sequencing of the *Drosophila* genome identified ~13,600 genes encoding 14,113 transcripts because of alternate splicing. In humans, 75,000–100,000 genes were expected on the basis of physical mapping; of these about 4000 may involve hereditary illness or cancer. The human gene number estimates in 2001 still varied from ~27,000 to ~150,000. In *Saccharomyces*, in the 5885 open reading frames, 140 genes encode rRNA, 40 snRNA and 270 tRNA. About 11% of the total protein produced by the yeast cells (proteome) has a metabolic function; 3% each is involved in DNA replication and energy production; 7% is dedicated to transcription; 6% to translation; and 3% (200) constitutes different transcription factors. About 7% is concerned with transporting molecules and about 4% constitutes structural proteins. Many proteins are involved with membranes. In *Caenorhabditis*, 19,099 protein-coding genes are predicted on the basis of the sequencing of the genome. The minimal essential gene number has also been estimated by comparing presumably identical genes in the smallest free-living cells *Mycoplasma genitalium* and *Haemophilus influenzae*, both completely sequenced. Insertional inactivation mutagenesis indicated the minimal number to be ~265–300. In *Caenorhabditis elegans*, about 20 times more genes are indispensable for survival. In higher organisms, the number of open reading frames may be larger than the number of essential genes (Rédei, 2002).

The gene number may not accurately reflect the functional complexity of a genome or organism because the combinatorial arrangement of proteins may generate great diversity and specificity. A synopsis of how these genes function would be most rewarding if one were able to present it even as a bird's-eye view. The most simplistic views are in the daily newspapers.

This sweeping and selective overview has missed out much important historical development. Fortunately, quantitative and population genetics have been better dealt with by many speakers than I ever could have attempted. I shall deal briefly with an area with which I was especially involved and which may have great significance for the future from the viewpoint of quantitative analysis.

Transformation

Transformation goes back to the late 1920s but it became practical with eukaryotes in the late 1970s and the early 1980s. By the mid-1980s, I was fortunate to be associated with researchers at the Max-Planck-Institut, Cologne, Germany. This effort resulted in the application of *in vivo* transcriptional gene-fusion technology to plants (Fig. 1.5).

In a similar manner, *in vivo* translational gene-fusion vectors can also be constructed in which there are no stop codons in front of the reporter gene and the translation initiation codon is removed, so the plant host protein and the reporter gene fusion would be facilitated. Obviously transformation provides unique opportunities to manipulate the genome and facilitates new insights into how in plants indigenous genes and foreign genes are regulated and expressed.

The Future of Genetics

It is customary to finish presentations with some predictions. Why I am shying away from general forecasts may be justified by a few more quotes.

Erwin Chargaff, the discoverer of the Chargaff rule, which was one of the cornerstones for the construction of the Watson and Crick model of the double helix, stated in 1955, only 6 years before the nature of the genetic code had been revealed:

> I believe, however, that while the nucleic acids, owing to the enormous number of possible sequential isomers, could contain enough

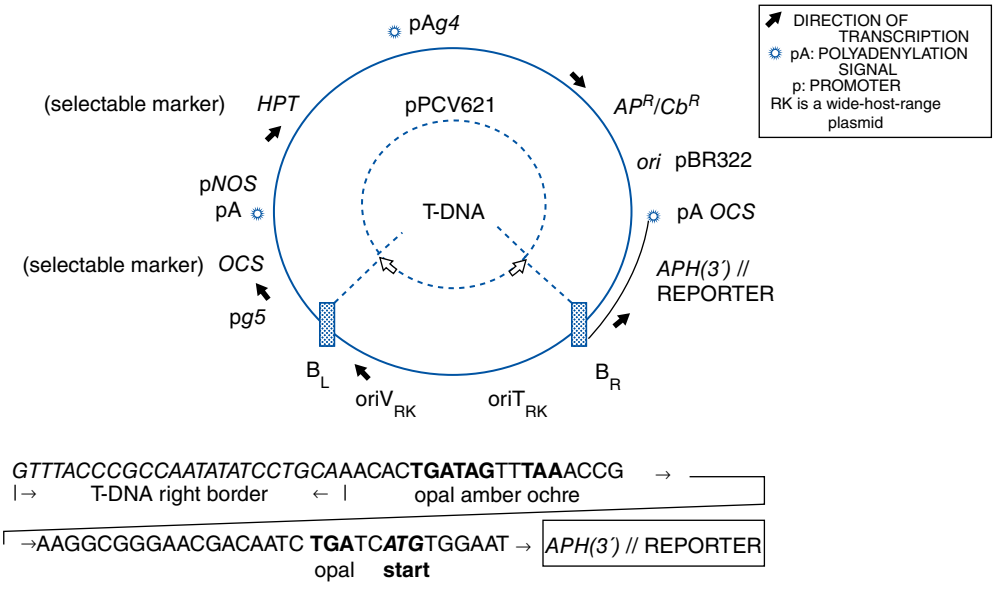


Fig. 1.5. The critical feature of this *in vivo* transcriptional gene-fusion vector is that the reporter *(aph(3′) II*, luciferase or *gus*) has no promoter and it is fused to the right border of the T-DNA. The structural gene of the reporter can be expressed only if it integrates behind a plant promoter that can provide the promoter function. In front of the structural gene here, there are four nonsense codons to prevent the fusion of the proteins with any plant peptide. (Based on oral communications by Dr Csaba Koncz.)

> codescripts to provide a universe with information, attempts to break the communications code of the cell are doomed to failure at the present very incomplete stage of our knowledge. Unless we are able to separate and to discriminate, we may find ourselves in the position of a man who taps all the wires of a telephone system simultaneously. It is, moreover, my impression that the present search for templates, in its extreme mechanomorphism, may well look childish in the future and that it may be wrong to consider the mechanisms through which inheritable characteristics are transmitted or those through which the cell repeats itself as proceeding in one direction only.

J.D. Watson's letter to Max Delbrück on 22 March 1953 sounds quite surprising today:

> I have a rather strange feeling about our DNA structure. If it is correct, we should obviously follow it up at a rapid rate. On the other hand, it will, at the same time, be difficult to avoid the desire to forget completely about nucleic acid and to concentrate on other aspects of life.
>
> (Judd, 1979)

The only remark I care to make is that I feel very assured that the role of quantitative genetics in basic biology and applied sciences will increase.

References

Aldred, C. (1961) *The Egyptians. Ancient People and Places*. Thames and Hudson, New York.

Aristotle (384–322 BC) *Generation of Animals*, Book IV, Pt I.

Auerbach, C. and Robson, J.M. (1944) Production of mutations by allylisothiocyanate. *Nature* (*London*) 154, 81.

Bateson, W. (1926) Segregation: being the Joseph Leidy Memorial Lecture at the University of Pennsylvania, 1922. *Journal of Genetics* 16, 201–235.

Baur, E. (1909) Das Wesen und die Erblichkeitsverhältnisse der 'varietates albomarginatae hort' von *Pelargonium zonale. Z. Ind. Abst. Vererb.-Lehre* 1, 330–351.

Belling, J. (1928) The ultimate chromomeres of *Lilium* and *Aloe* with regard to the numbers of genes. *Univ. Calif. Publ. Bot.* 14, 307–318.
Blixt, S. (1975) Why didn't Gregor Mendel find linkage? *Nature* (*London*) 256, 206.
Borek, E. (1965) *The Code of Life*. Columbia University Press, New York.
Borsook, H. (1956) Informal remarks 'by way of a summary'. *Journal of Cellular and Comparative Physiology* 47 (Suppl. 1), 283–286.
Carlson, E.A. (1966) *The Gene: a Critical History*. Saunders, Philadelphia, Pennsylvania.
Caspersson, T., Farber, S., Foley, G.E., Kudynowski, J., Modest, E.J., Simonsson, E., Wagh, U. and Zech, L. (1968) Chemical differentiation along metaphase chromosomes. *Experimental Cell Research* 49, 219–222.
Chargaff, E. (1955) On the chemistry and function of nucleoproteins and nucleic acids. *Istituto Lombardo* (*Rend. Sci.*) 89, 101–115.
Correns, C. (1900) G. Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Ber. Dtsch. Bot. Ges.* 18, 158–168.
Correns, C. (1902) Über den Modus und den Zeitpunkt der Spaltung der Anlagen bei den Bastarden vom Erbsen-Typus. *Bot. Zeitung* 60(2), 65–82.
Correns, C. (1909) Verebungsversuche mit blaß(gelb)grünen und buntblättrigen Sippen bei Mirabilis, Urtica und Linaria annua. *Z. Ind. Abst. Vererb.-Lehre* 1, 291–329.
de Vries, H. (1900) Sur la loi de disjonction des hybrides. *Comptes Rendus Académie de Science Paris* 130, 845–847.
de Winiwarter, H. (1912) Études sur la spermatogenèse humaine. *Arch. Biol.* 27, 91–189 + VIII plates.
Doncaster, L. and Raynor, G.H. (1906) Breeding experiments with Lepidoptera. *Proceedings of the Zoological Society London* 1, 125–133.
Elseth, G.D. and Baumgardner, K.D. (1984) *Genetics*. Addison-Wesley, Reading.
Fisher, G.H. (1968) Ambiguity of form: old and new. *Perception and Psychophysics* 4, 189–192.
Fisher, R.A. (1936) Has Mendel's work been rediscovered? *Annals of Science* 1, 115–137.
Focke, W.O. (1881) *Die Pflanzenmischlinge. Eine Beitrag zur Biologie der Gewächse*. Borntrager, Berlin.
Geitler, L. (1938) *Chromosomenbau*. Bornträger, Berlin.
Gray Lab Internet Glossary (2001) www.graylab.ac.uk/omd/
Haacke, W. (1893) Die Träger der Vererbung. *Biol. Zbl.* 13, 525–542.
Haartman, J.J. (1751) *Plantae Hybridae*. Uppsala, Sweden.
Henking, H. (1891) Über Spermatogenese bei *Pyrrhocoris apetrus*. *Zeitschr. Wissensch. Zool.* 51, 685–736.
Janssens, F.A. (1909) La théorie de la chiasmatypie. Nouvelle interprétations des cinèses de maturation. *Cellule* 25, 389–411.
Johannsen, W. (1909) *Elemente der exakten Erblichkeitslehre*. Fischer, Jena.
Judson, H.F. (1979) *The Eighth Day of Creation*. Simon and Schuster, New York.
Klug, W.S. and Cummings, M.R. (1983) *Concepts of Genetics*. Merrill, Columbus.
Koncz, C., Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Reiss, B., Rédei, G.P. and Schell, J. (1990) Isolation of a gene encoding a novel chloroplast protein by T-DNA tagging in *Arabidopsis thaliana*. *EMBO Journal* 9, 1337–1346.
Křiženecký, J. and Němec, B. (1965) *Fundamenta Genetica*. Czechoslovak Academy of Sciences, Prague.
Lander, E.S. with the International Human Genome Sequencing Consortium (2001) Initial Sequencing and Analysis of the Human Genome. Nature (London) 409, 860–921.
Lindegren, C.C. (1932) The genetics of *Neurospora* II. Segregation of the sex factors in asci of *N. crassa*, *N. sitophila* and *N. tetrasperma*. *Bulletin of the Torrey Botany Club* 59, 119–138.
Medvedev, Z. (1969) *The Fall and Rise of T.D. Lysenko*. Columbia University Press, New York.
Mendel, G. (1866) Versuche über Pflanzenhybriden. *Verh. Naturforsch. Verein. Brünn* 4, 3–47.
Migula, W. (1897) *System der Bakterien. Handbuch der Morphologie, Entwicklungsgeschichte und Systematik der Bakterien*. Fischer, Jena.
Moraes, C.T. (1998) Characteristics of mitochondrial DNA disease. In: Singh, K.K. (ed.) *Mitochondrial DNA Mutations and Aging*, Disease and Cancer. Springer-Verlag, Berlin, pp. 167–184.
Morgan, T.H. (1909) What are 'factors' in Mendelian explanations? *American Breeders' Association Report* 5, 365–368.
Morgan, T.H. (1910) Sex limited inheritance in Drosophila. *Science* 32, 120–122.
Morgan, T.H. (1911) An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila. *Journal of Experimental Zoology* 11, 365–412.
Morgan, T.H. (1919) *The Physical Basis of Heredity*. Lippincott, Philadelphia, Pennsylvania.
Morgan, T.H. (1926) Genetics and physiology of development. *American Nature* 60, 490–515.

Muller, H.J. (1927) Artificial transmutation of the gene. *Science* 66, 84–87.
Muller, H.J. (1941) Induced mutations in *Drosophila*. Cold Spring Harbor *Symposium on Quantitative Biology* 9, 151–167.
Nägeli, C.W. (1867) Cited in Gregor Mendel's letter to Carl Nägeli on 18 April 1867, p. 5. *Genetics* 35 (Suppl.), 1950.
Nägeli, C.W. (1884) *Mechanisch-physiologische Theorie der Abstammungslehre*. Oldenburg, Munich.
Olby, R.C. (1966) *Origins of Mendelism*. Schocken Books, New York.
Pearson, K. (1904) On a generalized theory of alternative inheritance, with special reference to Mendel's law. *Philosophical Transactions of the Royal Society A* 203, 53–86.
Plate, L. (1913) *Vererbungslehre*. Engelmann, Leipzig.
Pliny, G., II (AD 23–79) *C. Plinii Secundi Naturalis Historia*. Edited by Detlefsen, D. Berolini, Apud Weidmannos, 1866–1882.
Preer, J.R., Jr (1963) Discussion. In: Burdett, W.J. (ed.) *Methodology in Basic Genetics*. Holden-Day, San Francisco, p. 374.
Provine, W.B. (1971) *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago.
Punnett, R.C. (1949) Early days of genetics. *Heredity* 4, 1–10.
Rédei, G.P. (1974) Steps in the evolution of genetic concepts. *Biol. Zbl*. 93, 385–424.
Rédei, G.P. (2002) *An Encyclopedia Dictionary of Genetics, Genomics and Proteomics*. John Wiley & Sons, New York.
Rédei, G.P., Acedo, G.N. and Sandhu, S.S. (1984) Mutation induction and detection in *Arabidopsis*. In: Chu, E.H.Y. and Generoso, W.M. (eds) *Mutation, Cancer, and Malformation*. Plenum, New York, pp. 285–313.
Roberts, H.F. (1965) *Plant Hybridization before Mendel*. Hafner, New York.
Russel, P.J. (1992) *Genetics*. Harper Collins, New York.
Speicher, M.R., Ballard, S.G. and Ward, D.C. (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12, 368–375.
Stadler, L.J. (1928) Mutations in barley induced by X-rays and radium. *Science* 68, 186–187.
Stevens, N.M. (1905) Studies in spermatogenesis with special reference to the 'accessory chromosome'. *Carnegie Institute Publications* 36, 1–32.
Stoneking, M. (2001) Single nucleotide polymorphisms. From the evolutionary past … *Nature (London)* 409, 821–822.
Strickberger, M.W. (1985) *Genetics*. Macmillan, New York.
Sturtevant, A.H. (1965) *A History of Genetics*. Harper & Row, New York.
Sutton, E.H. (ed.) (1959) *Genetics: Genetic Control of Protein Structure*. Josiah Macey Foundation, Madison Press, Madison, Wisconsin, pp. 160–161.
Sutton, W.S. (1903) The chromosomes in heredity. *Biological Bulletin* 4, 231–248.
Suzuki, D.T., Griffiths, A.J.F. and Lewontin, R.C. (1976) *An Introduction to Genetic Analysis*. Freeman, San Francisco.
Sylvester, J. (1880) *The Complete Works*, Vol. I, p. 420. Edited by A.B. Grossart. Printed for private circulation, Edinburgh, UK.
Tijo, J.H. and Levan, A. (1956) The chromosome number of man. *Hereditas* 42, 1–6.
Udney Yule, G. (1907) On the theory of inheritance of quantitative compound characters on the basis of Mendel's laws. In: *Report of the Third International Conference on Genetics*. Spottiswood, London, pp. 140–142.
van Beneden, E. (1883) Recherches sur la maturation de l'oeuf et la fécondation. *Arch. Biol*. 4, 265–640.
von Nachtsheim, H. (1959) Chromosomenaberrationen beim Säuger und ihre Bedeutung für die Entstehung von Mißbildungen. *Naturwiss*. 46, 637–645.
Watson, J.D. (1953) Letter to Max Delbrück on 22 March 1953. Quoted after Judson, H.F. (1979) *The Eighth Day of Creation*. Simon and Schuster, New York, p. 229.
Weiling, F. (1966) Hat J.G. Mendel bei seinen Versuchen 'zu genau' arbeitet? – Der χ^2-Test und seine Bedeutung für die Beurteilung genetischer Spaltungsverhältnisse. *Züchter* 36, 359–365.
Wilson, E.B. (1905) The chromosomes in relation to the determination of sex in insects. *Science* 22, 500–502.
Woltereck, R. (1909) Weitere experimentelle Untersuchungen über Artveränderung, speziell über des Wesen quantitativer Unterschiede der Daphniden. *Verhandl. Dtsch. Zool. Ges.* 110–172.

Section I:

Quantitative Genetics: Plant Breeding, Bioinformatics, Genome Editing and G×E Interaction

2 Food and Health: The Role of Plant Breeding

Salvatore Ceccarelli*
Rete Semi Rurali, Scandicci, Italy

Introduction

The current situation about food security and nutrition worldwide is well summarized in a report published in October 2018 (FAO, IFAD, UNICEF, WFP and WHO, 2018), which signals that the absolute number of undernourished people, i.e. those facing chronic food deprivation, has increased to nearly 821 million in 2017, from around 804 million in 2016. These are levels from almost a decade ago. The human cost of our food system has been estimated at 1 billion people hungry, nearly 2 billion people who are eating too much of the wrong foods and 11 million people who die prematurely because of unhealthy diets, while agricultural production is neither resilient nor sustainable (Lucas and Horton, 2019).

The relationship between diet and health is through the microbiota – namely, the complex of bacteria, viruses, fungi, yeast and protozoa, which are in our intestine (sometimes called microbiome, which actually refers to the genes of the microbiota). A change in the diet changes its composition in just 24 h; it takes 48 h, after changing the diet again, before the microbiota returns to the initial conditions (Singh *et al.*, 2017). The microbiota are associated with our immune system and then with the possibility of contracting or not contracting inflammatory diseases (Khamsi, 2015).

The microbiota also appears to be involved in several neuropsychiatric disorders, such as depression, schizophrenia, autism, anxiety, stress response (Hoban *et al.*, 2016) and quality of life (Valles-Colomer *et al.*, 2019). This is likely attributable to the damage that inflammatory processes cause to myelin, the sheath surrounding the neurons, thus altering the normal transmission of nerve impulses.

Given the important roles of the microbiota on one hand, and the fact that these are so strongly and rapidly influenced by diet on the other hand, it is understandable that there have been many studies on the effects of various diets (Western, omnivorous, Mediterranean, vegetarian, vegan, etc.) on the composition and the diversity of the microbiota itself (Singh *et al.*, 2017). Recent results demonstrate that gut microbiota composition is shaped predominantly by environmental factors (diet and lifestyle) and that the microbiota are not significantly associated with genetic ancestry (Rothschild *et al.*, 2018).

The diet also links environmental and human health. Rising incomes and urbanization are driving a global dietary transition, in which traditional diets are being replaced by diets higher in refined sugars, refined fats, oils and meats. By 2050, these dietary trends, if unchecked, would be a major contributor to an estimated 80% increase

* Email: ceccarelli.salvatore83@gmail.com

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

in global agricultural greenhouse gas emissions from food production and to global land clearing. Moreover, these dietary shifts are greatly increasing the incidence of type II diabetes, coronary heart disease and other chronic, non-communicable diseases that lower global life expectancies (Tilman and Clark, 2014). The opinions of the nutritionists examining the effects of various diets do not always agree, but what all nutritionists seem to agree on is that diet diversity is of paramount importance for having a healthy microbiota (Heiman and Greenway, 2016).

Diet Diversity and Crop Uniformity

How can we have a diversified (healthy) diet if 60% of our calories come from just three crops, namely wheat, rice and maize (Thrupp, 2000)? Incidentally, these three crops are far less nutritious than crops such as barley (Grando and Gomez Macpherson, 2005), and millets and sorghum (Dwivedi *et al.*, 2011; Boncompagni *et al.*, 2018). Millets and sorghum need less water than maize, rice and wheat, which use nearly 50% of all the water used for crop irrigation.

Crucial questions are: How do we diversify our food if almost all the food we eat is produced from varieties that, to be legally marketed, i.e. for their products to be legally found in supermarkets, must be registered in a catalogue called the varietal register, and that to be registered, should be uniform, stable and distinct? If our health depends on the diversity and composition of the microbiota, which in turn depends on the diversity of the diet, how can we have a diversified diet if the agriculture that produces our food is based on uniformity?

Between the need to diversify our diet and the uniformity imposed by law on seed, and thus on crops, there is an obvious contradiction. In addition, there is also an obvious contradiction between uniformity and stability on the one hand, and the need to adapt crops to climate change on the other hand.

Already in 1950, Sir Otto Frankel warned, 'From the early days of plant breeding, uniformity has been sought after with great determination. For this, there are many reasons: technical, commercial, historical, psychological, and aesthetic,' and, 'the concept of purity has not only been carried to unnecessary length but that it may be inimical to the attainment of highest production' (Frankel, 1950, p. 97). Notice the absence in the quotation of terms such as scientific or biological.

Biodiversity, and particularly agrobiodiversity, is of critical importance for food security (Zimmerer and de Haan, 2017). Agrobiodiversity has been shown to be highly beneficial, particularly in restricting the development of diseases (Zhu *et al.*, 2000; Döring *et al.*, 2011).

In addition to the increased uniformity of the varieties that we grow, plant breeding has also contributed to the decrease in the number of crops, with only about 30 plant species supplying 95% of the global demand for food (FAO, 2010). The four biggest staple crops (wheat, rice, maize and potato) account for the lion's share of food production (Esquinas-Alcázar, 2005). This is reflected in the investment in research, which in the global public sector is still focused mostly on rice, wheat and maize, with 45% of private-sector investment in agricultural research being in maize (Haddad *et al.*, 2016).

Seeds, Food and Health

Alongside a food oligopoly, there is also a seed oligopoly; in fact, about 55% (2016 data) of the world seed market, which is worth billions of dollars, is in the hands of five large multinational corporations (Bonny, 2017), up from only 10% in 1985. With the recent mergers, the number of dominant corporations has been further reduced to four (Clapp, 2017). Some of those corporations simultaneously control another multi-billion-dollar market, that of pesticides (i.e. herbicides, insecticides and fungicides). The application of Big Data in the agri-food sector in a variety of information platforms, which links physical data with seed, pesticides and farm equipment, will exacerbate the 'productivist' model of agriculture with overproduction of inexpensive, low-nutrient food and will increase the inequities between farmers and large corporations (Bronson and Knezevic, 2016).

This is quite worrying, because it has been established beyond reasonable doubt that there is a close relationship between exposure to pesticides and the rise of chronic diseases. This has been shown in the case of different types of cancer, diabetes, neurodegenerative disorders, such as Parkinson's, Alzheimer's and ALS (amyotrophic

lateral sclerosis), birth defects and reproductive disorders (Mostafalou and Abdollahi, 2013; von Ehrenstein *et al.*, 2019).

The breeding philosophy by which selection must be conducted in an optimum agronomic environment to maximize genetic differences and hence selection gains, is likely to be at the root of the decline in cultivated diversity, but most importantly is at the root of the dependence of modern varieties on chemical inputs. The latter are in fact used to create an optimum agronomic environment within the research stations, and when used by farmers, tend to smooth the differences between locations, thus making possible for a relatively small number of varieties to thrive well across large geographical areas (Ceccarelli, 1989).

On one hand, this has led to a large increase in agricultural production, but on the other hand has already transgressed some planetary boundaries such as genetic diversity and biogeochemical flows, namely nitrogen and phosphorus cycles (Campbell *et al.*, 2017). Nitrogen (N) in particular, is used in excessive amounts and about half of applied N is lost through leaching (16%), soil erosion (15%) and gas emission (14%) (Bodirsky *et al.*, 2012). These effects add on to those described earlier on world hunger and human health.

How Consumers Defend Themselves

A solution that more and more consumers have chosen in recent times has been to turn to products of organic agriculture, which, while on one hand, guarantees healthy food; on the other hand, is not free from criticism. In fact, although organic farming can be the solution for its multiple benefits (Reganold and Wachter, 2016), it is often criticized for producing expensive food and is dubbed as a method for starving 2 billion people (Connor, 2008). The first criticism can be easily answered: in fact, the real problem is not that organic food is too expensive but that the real costs that consumers must pay in the supermarkets for the non-organic food are hidden. These are the negative effects of the food industry driving policy of producing food at low cost, on the environment (soils, water and air), no matter what the ultimate cost is (Sukhdev *et al.*, 2016). Furthermore, upscaling the area under organic agriculture can directly contribute to many of the 17 Sustainable Development Goals, such as poverty, zero hunger, good health and wellbeing, clean water and sanitation, responsible consumption and production, climate action and life on land (Eyhorn *et al.*, 2019). There are also effects of non-organic foods on our health: the global cost of diabetes in 2015 was estimated at US$ 1.31 trillion (Bommer *et al.*, 2017). In Africa, 35 million people – twice the number at present – will be affected by diabetes in the next 20 years (Jaffar, 2016).

The health benefits of organic food, particularly in relation to different types of cancer, has been shown recently by one of the first large-scale studies on the relationship between organic food consumption and cancer risks involving 68,946 French adults: the study found a 25% lower overall cancer risk among those consuming organic food most regularly compared to those who reported very little or no consumption of organic food (Baudry *et al.*, 2018).

The second criticism, namely that the production from organic farming is lower (on average, between 8% and 25%) than that from conventional agriculture, depending on the crop and the way in which organic farming is practised, is used more often than the first one to argue that with organic farming, many more people would suffer from hunger. People ask, what is the point of questioning such an agricultural system (conventional farming) when we need to increase agricultural production by 70% or even 100% by 2050?

There is evidence that yields under organic conditions are generally higher than yields under conventional conditions in years affected by drought (Lotter *et al.*, 2003); therefore, one of the consequences of the effects of climate change could well be reducing the yield gap between conventional and organic agriculture.

Recently, the mathematical models on which this second critique is based have been questioned. First, when we discuss global agricultural production, we tend to forget the enormous quantities of food that are wasted annually, as much as 1.3 billion tonnes, equal to 30% of agricultural production (FAO, 2011). In addition, it is estimated that, globally, we produce about 4600 kilocalories per person per day; of which 1400 are lost after harvest, during distribution and consumption. The 3200 kilocalories left are

still almost 1000 more than the 2360 kilocalories per person per day that, according to the World Health Organization, are sufficient for a healthy life (Smil, 2000). Thus, the farmers of the world already produce around 50% more food than we need and, arguably, more than we will ever need.

Therefore, the hypothesis is that the need to increase agricultural production by 70% or even 100% by 2050 has been questioned as an overestimate. It is one deliberately used by institutions and individuals with a prior existing set of ideological commitments regarding the problem of food security (Tomlinson, 2013). In other words, these data are used to justify the need for pesticides and of biotechnologies, such as genetically modified organisms (GMOs), and, more recently, the gene editing that has aroused so much controversy in the European Union (EU).

The main problem with GMOs is that, because of the Fundamental Theorem of Natural Selection (Shaw and Shaw, 2014), they, at best, represent a temporary solution to the problem they intend to solve because weeds and pests evolve resistance as shown by Gray *et al.* (2009), Gassmann *et al.* (2011), Fisher (2012), Hagenbucher *et al.* (2013a) and McDonald and Stukenbrock (2016). They may also affect non-target species either causing their spread (Hagenbucher *et al.*, 2013b; Lu *et al.*, 2013) or causing the death of non-target insects (Hilbeck *et al.*, 2019).

The controversy about gene editing has focused on whether a gene-edited organism should be considered a GMO or not. Actually, the real issues with gene editing are: (i) the technique seems to cause uncontrolled mutations in other genomic regions (Kosicki *et al.*, 2018; *The Lancet*, 2018); (ii) even if these problems are overcome, the resulting organism, although technically not a GMO, will suffer, like a GMO, of the non-durability of single gene solutions to pest resistance; and (iii) several important agronomic traits, and even those related to adaptation, are quantitative in nature, and are controlled by quantitative trait loci (QTL) spread throughout the genome: it remains to be seen how these traits can be manipulated through gene editing.

What Can Plant Breeding Do?

Food derives from seeds and therefore the primary cause of the health problems afflicting the world today as well as those affecting the planet (Steffen *et al.*, 2015) is the way in which the seeds are produced. In addition, since the seeds are produced by plant breeding, the solution may require reconsidering how plant breeding is done to shift from 'cultivating uniformity' to 'cultivating diversity'.

Today, much of the 'institutional' plant breeding, and not just the private plant breeding, has as its objective industrial agriculture (the only way, according to some, we will be able to feed the world, while estimates suggest that family farms produce more than 80% of the world's food in value terms) (FAO, 2014). Institutional plant breeding is based on the selection, in one or few research centres, of uniform varieties – to oblige the seed laws mentioned earlier – and able to maximize production with the support of fertilizers and pesticides.

Plant breeding programmes specifically addressing organic farming are very limited in number (Campanelli *et al.*, 2015). Therefore, one of the reasons for the yield gap between conventional agriculture and organic farming is that in the latter, lacking specifically adapted varieties, the same varieties are grown that are bred for conventional agriculture. Those varieties obviously find themselves in different agronomic conditions from the ones under which they were selected, and therefore produce less.

The main difference between conventional (industrial) agriculture and organic agriculture is not only in the nature of the inputs that are allowed in each of the two systems, but, from a plant breeding viewpoint, it is in the way the respective breeding programmes are organized. In the case of industrial agriculture, being predominantly based on uniformity, a centralized plant breeding programme is well suited particularly if supported by a centralized seed system.

On the contrary, because of its nature, organic agriculture is much more location specific than industrial agriculture, and therefore needs to be supported by both a decentralized plant breeding programme and a decentralized seed system based on a plurality of small seed companies.

The objective of 'cultivating diversity' can be achieved quickly and inexpensively with evolutionary plant breeding (Suneson, 1956; Ceccarelli, 2009), which consists of cultivating mixtures or populations.

The science of evolutionary plant breeding (EPB), even if the material developed by this

method was initially called composite cross population (CCP) or Bulks, started with the work of Harlan and Martini (1929) in Washington DC. In the following years, papers by Harlan and Martini (1938), Suneson and Wiebe (1942), Suneson (1956), Allard and Hansche (1964), Patel *et al.* (1987), Ibrahim and Barret (1991), Soliman and Allard (1991), Wolfe (1991) and Wolfe *et al.*, (1992) showed that evolutionary populations do evolve, increasing their yield and their temporal stability, becoming more resistant to diseases, becoming useful source material to develop high-yielding lines and adapting their phenology to the location where they evolve (Goldringer *et al.*, 2006). They tend to perform better than uniform varieties in years affected by drought (Danquah and Barrett, 2002) and they can actually combine high yield and high yield stability (Raggi *et al.*, 2016a,b, 2017). More recently, data from baking tests showed that CCPs specifically created for good baking quality are as good in terms of baking volume or better (protein content and Hagberg falling number) than modern elite wheat varieties (Brumlop *et al.*, 2017).

Once a mixture or a population is planted, it is left to evolve as a crop, i.e. is planted and harvested year after year, using part of the harvest as seed for the next season. At the same time, the farmer, possibly assisted by a breeder in what becomes evolutionary participatory plant breeding, can use the evolving population as a source to select the best plants. Thanks to the natural hybridization that always occurs between plants (more frequent in cross-pollinated than in self-pollinated plants) and the effect of natural selection, the seed that is harvested is genetically different from the one that was planted. In other words, the genetic composition of the populations (including those derived from an original mixture) changes because they evolve continuously (this is why they are called 'evolutionary') and therefore the farmers have the opportunity to adapt the crops to their soil, to their climate and to the particular way in which each of them practices agriculture, including organic farming. It is a biological form of precision agriculture.

Summary and Conclusions

Our health is associated with our diet through the microbiota, which, depending on their composition and diversity, affects our immune system and hence both our physical and mental health. There is a contradiction between the need for diet diversity and crop uniformity, which is the main feature of industrial agriculture. There is also a contradiction between crop uniformity and the need to adapt crops to both short- and long-term climate change and associated changes in the spectrum of pests and diseases. The role of plant breeding is to provide scientific support to cultivate diversity through an emphasis on specific spatial adaptation and wide temporal adaptation through decentralized selection, and with the use of evolutionary plant breeding methods. These methods have received official recognition in Europe for some important food crops, such as wheat, barley, oats and maize, which allows the seed of evolutionary populations to be legally commercialized. Reductionist approaches, such as those used to produce GMOs, gene editing and genomic selection, do not appear adequate to tackle the complexity of the challenges ahead.

A mixture or a population appears to be an integrated solution to several problems: it evolves gradually, adapting to every single microenvironment within the same farm, guarantees the farmer a stable production despite the climatic variations from one year to the next, is able to control weeds, insects and diseases without the use of chemicals, and therefore, unlike GMOs and others single-gene quick fixes, offers a durable solution to the problem of resistance. Furthermore, they gradually adapt to long-term climate change, ensure income to the farmer both as seed and as grain, guarantee the consumer a healthy product, make the farmer the owner of his own seed and put the farmer in a position to manage his own future. Eventually, populations and mixtures can offer an immediate solution to every new problem that may appear, thanks to their diversity.

In the case of wheat, maize, rice and oats, the commercialization of seed of evolutionary populations in Europe has been made legal by the Commission Implementing Decision of 18 March 2014 pursuant to Council Directive 66/402/EEC, which was extended to 2021 in October 2018. With the Council Directive it is presently possible to market experimentally heterogeneous materials of the four different cereals. Consequently, only in Italy, eight populations

(three bread wheat, four durum wheat and one barley) have been authorized and their seed is being commercialized. Because of their ability to control pests, these populations are grown mostly by organic farmers.

The two evolutionary populations of bread wheat and durum wheat were assembled in 2009. Nine years later, the bread and the pasta produced with them were already on the shelves of a number of shops.

References

Allard, R.W. and Hansche, P.E. (1964) Some parameters of population variability and their implications in plant breeding. In: Norman, A. (ed.) *Advances in Agronomy*. Academic Press, New York, pp. 281–325.

Baudry, J., Assmann, K.E., Touvier, M., Allès, B., Seconda, L., *et al*. (2018) Association of frequency of organic food consumption with cancer risk: Findings from the NutriNet-Santé prospective cohort study. *JAMA Internal Medicine* 178(12), 1597–1606.

Bodirsky, B.L., Popp, Weindl, A.I., Dietrich, J.P., Rolinski, S., *et al*. (2012) N_2O emissions from the global agricultural nitrogen cycle – current state and future scenarios. *Biogeosciences* 9, 4169–4197.

Bommer, C., Heesemann, E.V., Sagalova, J., Manne-Goehler, R., Atun, T., *et al*. (2017) The global economic burden of diabetes in adults aged 20–79 years: A cost-of-illness study. *The Lancet Diabetes & Endocrinology* 5 (6), 423–430, DOI: 10.1016/S2213-8587(17)30097-9.

Boncompagni, E., Orozco-Arroyo, G., Cominelli, E., Gangashetty, P.I., Grando, S., *et al*. (2018) Antinutritional factors in pearl millet grains: Phytate and goitrogens content variability and molecular characterization of genes involved in their pathways. *PLoS ONE* 13(6), e0198394, DOI: 10.1371/journal.pone.0198394.

Bonny, S. (2017) Corporate concentration and technological change in the global seed industry. *Sustainability* 9, 1632.

Bronson, K. and Knezevic, I. (2016) Big Data in food and agriculture. *Big Data & Society*, DOI: 10.1177/2053951716648174.

Brumlop, S., Pfeiffer, T. and Finckh, M.R. (2017) Evolutionary effects on morphology and agronomic performance of three winter wheat composite cross populations maintained for six years under organic and conventional conditions. *Organic Farming* 3(1), 34–50, DOI: 10.12924/of2017.03010034.

Campanelli, G., Acciarri, N., Campion, B., Delvecchio, S., Leteo, F., *et al*. (2015) Participatory tomato breeding for organic conditions in Italy. *Euphytica* 204(1), 179–197.

Campbell, B.M., Beare, D.J., Bennett, E.M., Hall-Spencer, J.M., Ingram, J.S.I., *et al*. (2017) Agriculture production as a major driver of the earth system exceeding planetary boundaries. *Ecology and Society* 22(4), 8.

Ceccarelli, S. (1989) Wide adaptation. How wide? *Euphytica* 40, 197–205.

Ceccarelli, S. (2009) Evolution, plant breeding and biodiversity. *Journal of Agriculture and Environment for International Development* 103(1/2), 131–145.

Clapp, J. (2017) Bigger is not always better: The drivers and implications of the recent agribusiness megamergers. Waterloo ON: Global Food Politics Group, University of Waterloo. March 2017.

Connor, D.J. (2008) Organic agriculture cannot feed the world. *Field Crops Research* 106(2), 187–190.

Danquah, E.Y. and Barrett, J.A. (2002) Grain yield in composite cross five of barley: Effects of natural selection. *Journal of Agricultural Science* 138, 171–176.

Döring, T.D., Knapp, S., Kovacs, G., Murphy, K. and Wolfe, M.S. (2011) Evolutionary plant breeding in cereals – into a new era. *Sustainability* 3, 1944–1971.

Dwivedi, S., Upadhyaya, H., Senthilvel, S., Hash, C., Fukunaga, K., *et al*. (2011) Millets: Genetic and genomic resources. In: Janick, J (ed.) *Plant Breeding Reviews*, Volume 35, Wiley-Blackwell, Hoboken, New Jersey, pp. 247–377.

Esquinas-Alcázar, J. (2005) Protecting crop genetic diversity for food security: Political, ethical and technical challenges. *Nature Reviews Genetics* 6, 946–953.

Eyhorn, F., Muller, A., Reganold, J.P., Frison, E., Herren, H.R., *et al*. (2019) Sustainability in global agriculture driven by organic farming. *Nature Sustainability* 2, 253–255.

FAO (2010) *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome. Available at: www.fao.org/docrep/013/i1500e/i1500e.pdf (accessed 20 August 2015).

FAO (2011) *Global Food Losses and Food Waste – Extent, Causes and Prevention*. FAO, Rome.

FAO (2014) *The State of Food and Agriculture 2014. Innovation in Family Farming*. FAO, Rome.

FAO, IFAD, UNICEF, WFP and WHO (2018) *The State of Food Security and Nutrition in the World 2018. Building Climate Resilience for Food Security and Nutrition*. FAO, Rome.
Fisher, M. (2012) Many little hammers: Fighting weed resistance with diversified management. *CSA News*, September 2012, 4–10.
Frankel, O.H. (1950) The development and maintenance of superior genetic stocks. *Heredity* 4, 89–102.
Gassmann, A.J., Petzold-Maxwell, J.L., Keweshan, R.S. and Dunbar, M.V (2011) Field-evolved resistance to Bt maize by western corn rootworm. *PLoS ONE* 6(7), e22629.
Goldringer, I., Prouin, C., Rousset, M., Galic, N. and Bonnin, I. (2006) Rapid differentiation of experimental populations of wheat for heading time in response to local climatic conditions. *Annals of Botany* 98(4), 805–817.
Grando, S. and Gomez Macpherson, H. (eds) (2005) Food barley: Importance, uses and local knowledge. *Proceedings of the International Workshop on Food Barley Improvement*, Hammamet, Tunisia, 14–17 January 2002, ICARDA, Aleppo, Syria, 156 pp.
Gray, M.E., Sappington, T.W., Miller, N.J., Moeser, J. and Bohn, M.O. (2009) Adaptation and invasiveness of western corn rootworm: Intensifying research on a worsening pest. *Annual Review of Entomology* 54, 303–321.
Haddad, L., Hawkes, C., Webb, P., Thomas, S., Beddington, J. (2016) A new global research agenda for food. *Nature* 540, 30–32.
Hagenbucher, S., Olson, D.M., Ruberson, J.R., Wäckers, F.L. and Romeis, J. (2013a) Resistance mechanisms against arthropod herbivores in cotton and their interactions with natural enemies. *Critical Reviews in Plant Sciences* 32, 458–482.
Hagenbucher, S., Wäckers, F.L., Wettstein, F.E., Olson, D.M., Ruberson, J.R., *et al*. (2013b) Pest trade-offs in technology: Reduced damage by caterpillars in Bt cotton benefits aphids. *Proceedings of the Royal Society B Biological Sciences* 280, 20130042.
Harlan, H.V. and Martini, M.L. (1929) A composite hybrid mixture. *Journal of American Society of Agronomy* 21, 487–490.
Harlan, H.V. and Martini, M.L. (1938) The effect of natural selection in a mixture of barley varieties. *Journal of Agricultural Research* 57(3), 189–199.
Heiman, M.L. and Greenway, F.L. (2016) A healthy gastrointestinal microbiome is dependent on dietary diversity. *Molecular Metabolism* 5(5), 317–320.
Hilbeck, A., Defarge, N., Bøhn, T., Krautter, M., Conradin, C., *et al*. (2019) Impact of antibiotics on efficacy of cry toxins produced in two different genetically modified Bt maize varieties in two lepidopteran herbivore species, *Ostrinia nubilalis* and *Spodoptera littoralis*. *Toxins* 10(12), 489, DOI: 10.3390/toxins10120489.
Hoban, A.E., Stilling, R.M., Ryan, F.J., Shanahan, F., Dinan, T.G., *et al*. (2016) Regulation of prefrontal cortex myelination by the microbiota. *Translational Psychiatry* 6, e774.
Ibrahim, K.M. and Barret, J.A. (1991) Evolution of mildew resistance in a hybrid bulk population of barley. *Heredity* 67, 247–256.
Jaffar, S. (2016) Diabetes and other non-communicable diseases in Africa, a potential disaster in the waiting. *The Lancet* 4(11), 875–877.
Khamsi, R. (2015) A gut feeling about immunity. *Nature Medicine* 21, 674–676.
Kosicki, M., Tomberg, K. and Bradley, A. (2018) Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 36(8), 765–771.
Lotter, D.W., Seidel, R. and Liebhardt, W. (2003) The performance of organic and conventional cropping systems in an extreme climate year. *American Journal of Alternative Agriculture* 18, 146–154.
Lu, Y., Wu, K., Jiang, Y., Xia, B., Li, P., *et al*. (2013) Mirid bug outbreaks in multiple crops correlated with wide-scale adoption of Bt cotton in China. *Science* 328, 1151–1154.
Lucas, T. and Horton, R. (2019) The 21st-century great food transformation. *The Lancet* 393(10170) 386–387.
McDonald, B.A. and Stukenbrock, E.H. (2016) Rapid emergence of pathogens in agro-ecosystems: Global threats to agricultural sustainability and food security. *Philosophical Transactions of the Royal Society B Biological Sciences* 371, 20160026.
Mostafalou, S. and Abdollahi, M. (2013) Pesticides and human chronic diseases: evidences, mechanisms, and perspectives. *Toxicology and Applied Pharmacology* 268(2), 157–77.
Patel, J.D., Reinbergs, E., Mather, D.E., Choo, T.M. and Sterling, J.D. (1987) Natural selection in a double-haploid mixture and a composite cross of barley. *Crop Science* 27, 474–479.
Raggi, L., Ceccarelli, S. and Negri, V. (2016a) Evolution of a barley composite cross-derived population: An insight gained by molecular markers. *The Journal of Agricultural Science* 154, 23–39.

Raggi, L., Negri, V. and Ceccarelli, S. (2016b) Morphological diversity in a barley composite cross derived population evolved under low-input conditions and its relationship with molecular diversity: Indications for breeding. *The Journal of Agricultural Science* 154, 943–959.
Raggi, L., Ciancaleoni, S., Torricelli, R., Terzi, V., Ceccarelli, S., *et al*. (2017) Evolutionary breeding for sustainable agriculture: Selection and multi-environment evaluation of barley populations and lines. *Field Crops Research* 204, 76–88.
Reganold, J.P. and Wachter, J.M. (2016) Organic agriculture in the twenty-first century. *Nature Plants* 2, 1–8, DOI: 10.1038/nplants.2015.221.
Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., *et al*. (2018) Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555, 210–215.
Shaw, R.G. and Shaw, F.H. (2014) Quantitative genetic study of the adaptive process. *Heredity* 112, 13–20.
Singh, R.K., Chang, H.W., Yan, D., Lee, K.M., Ucmak, D., *et al*. (2017) Influence of diet on the gut microbiota and implications for human health. *Journal of Translational Medicine* 15(1), 73.
Smil, V. (2000) *Feeding the World: A Challenge for the Twenty-First Century*. MIT Press, Cambridge, Massachusetts.
Soliman, K. M. and Allard, R.W. (1991) Grain yield of composite cross populations of barley: Effects of natural selection. *Crop Science* 31, 705–708.
Steffen, W., Richardson, K., Rockstrom, J., Cornell, S.E., Fetzer, I., *et al*. (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347(6223), 1259855–1259855.
Sukhdev, P., May, P. and Müller, A. (2016) Fix food metrics. *Nature* 540, 33–34.
Suneson, C.A. (1956) An evolutionary plant breeding method. *Agronomy Journal* 48, 188–191.
Suneson, C.A. and Wiebe, G.A. (1942) Survival of barley and wheat varieties in mixtures. *Journal of the Agronomy Society of America* 34, 1052–1056.
THE LANCET (2018) Genome editing: Proceed with caution. *The Lancet* 392(10144), 253.
Thrupp, L.A. (2000) Linking agricultural biodiversity and food security: The valuable role of agrobiodiversity for sustainable agriculture. *International Affairs* 76, 265 –281.
Tilman, D. and Clark, M. (2014) Global diets link environmental sustainability and human health. *Nature* 515(7528), 518–522.
Tomlinson, I. (2013) Doubling food production to feed the 9 billion: A critical perspective on a key discourse of food security in the UK. *Journal of Rural Studies* 29, 81–90.
Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E.F., Wang, J., *et al*. (2019) The neuroactive potential of the human gut microbiota in quality of life and depression. *Nature Microbiology* 4, 623–632.
von Ehrenstein, O.S., Ling, C., Cui, X., Cockburn, M., Park, A.S., *et al*. (2019) Prenatal and infant exposure to ambient pesticides and autism spectrum disorder in children: Population based case-control study. *British Medical Journal* 2019, 364, l962.
Wolfe, M.S. (1991) Barley diseases: Maintaining the value of our varieties. In: Munck, L. (ed.) *Barley Genetics VI*, Volume 2. Munksgaard Int. Publishers, Copenhagen, Denmark, pp. 1055–1067.
Wolfe, M.S., Brändle, U., Koller, B., Limpert, E., McDermott, J.M., *et al*. (1992) Barley mildew in Europe: Population biology and host resistance. *Euphytica* 63, 125–139.
Zhu, Y., Chen, H., Fan, J., Wang, Y., Li, Y., *et al*. (2000) Genetic diversity and disease control in rice. *Nature* 406, 718–722.
Zimmerer, K.S. and de Haan, S. (2017) Agrobiodiversity and a sustainable food future. *Nature Plants* 3, 17047.

3 The Importance of Plant Pan-genomes in Breeding

Soodeh Tirnaz, David Edwards and Jacqueline Batley*
University of Western Australia, Perth, Australia

Introduction

The study of different aspects of genomes, such as structural variations, has been facilitated by advances in efficient and inexpensive genome sequencing methods and the exponential increase in the amount of data obtained for different plant species. It has been demonstrated that a single reference genome cannot represent a full picture of genetic diversity in a species. For example, comparison of four genomic regions between maize (*Zea mays* L.) inbred lines B73 and Mo17 showed that around 25% of the sequences were identified in a homologous location in one of the inbred lines but not in the other, and on average, only 50% of their sequences are shared (Brunner *et al.*, 2005). Similar differences between rice (*Oryza sativa* L.) genotypes have been reported (Yu *et al.*, 2005). These observations highlight that a single genome sequence is insufficient to represent the genomic diversity of a species. The concept of pan-genomes was initially proposed and used by Tettelin *et al.* (2005) following genomic comparison between six strains of *Streptococcus agalactiae*, a human bacterial pathogen (Tettelin *et al.*, 2005). Pan-genomes help to describe the genomic variation within a species, and can be split into the core genome containing genes common to all individuals, and a dispensable (or variable) genome consisting of partially shared DNA sequence elements (Tettelin *et al.*, 2005). Pan-genomes are widely used as the basis for studying presence/absence variation (PAV) of genomic structures, such as genes.

Plant Pan-genomes

Pan-genomics is being used to study genomic structural variations (SVs), including PAVs and copy number variations (CNVs) in plant genomes. PAVs refer to the sequences that are present in one genome and absent in another genome, and they reflect the genomic information of individuals within a species. CNVs refer to sequences present in different copy numbers between individuals of a species (Saxena *et al.*, 2014). Both CNVs and PAVs are common in plant species and play important roles in the evolution and selection of genes associated with agronomic traits, such as those for biotic and abiotic stresses (Dolatabadian *et al.*, 2017). There are several causes of gene content variation, though in plants this is mostly attributable to rounds of polyploidy, leading to gene redundancy and

* Email: jacqueline.batley@uwa.edu.au

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

differential gene loss in subsequent variations. Polyploidy can also lead to non-reciprocal homologous exchange between the genomes, with associated gene loss, which can have important agronomic implications (Hurgobin *et al.*, 2018).

There are a diverse range of tools and approaches for pan-genome construction, which are discussed by Golicz *et al.* (2016a). Two approaches have mainly been used in the construction of plant pan-genomes (see Table 3.1). In brief, the first approach is based on *de novo* genome assemblies, followed by individual annotation and comparison of the gene content (Tettelin *et al.*, 2005). This approach has the advantage that the variable genes are often placed within their chromosomal location, though it suffers from issues with variable assembly and annotation of the individual genomes, which can lead to more erroneously called PAVs than genuine PAVs (Bayer *et al.*, 2017). The second method is reference-based mapping and assembly of unmapped reads, which starts by mapping sequence reads to an existing reference, followed by assembly of unmapped reads using a metagenomics-based genome assembler (Golicz *et al.*, 2016a,b). This approach has the advantage that it can use large numbers, possibly thousands of individuals with relatively low coverage (> 10x) to identify relatively rare genes in populations, and the data can be applied to call PAVs and single nucleotide polymorphisms (SNPs) across these populations. A disadvantage of this approach though is that only a portion (around 40%) of the novel genes can be placed in a chromosomal context. Given the differences between the two approaches, they are highly complementary, and the most thorough pan-genome analysis of a species would apply a combination of both approaches. In addition to these pan-genome assembly methods, some researchers also undertake pan-transcriptome analysis using whole-transcriptome sequencing (RNA-seq) data of individuals (Hirsch *et al.*, 2014). This was previously considered to be a gene expression atlas and is valuable for the analysis of gene expression, however it is unsuitable for the assessment of gene content because of some genes only being expressed in specific tissues or environmental conditions.

Pan-genomes have been constructed for important crop species, including wheat, rice, maize, *Brassica napus* and *Brassica oleracea* (Table 10.1), as well as for an increasing number of minor crop species, such as sesame (Yu *et al.*, 2019). The first rice pan-genome was constructed by Yao *et al.* (2015) using 1483 rice accessions from japonica and indica subspecies.

Table 3.1. Pan-genome studies of plant species.

Species	Number of genotypes	Method	Reference
Oryza sativa	50	Reference-based mapping and draft assembly of unmapped reads	(Xu *et al.*, 2012)
O. sativa	66	*De novo* draft genome assemblies	(Zhao *et al.*, 2018)
O. sativa	1483	Reference-based mapping and draft assembly of unmapped reads	(Yao *et al.*, 2015)
O. sativa	3010	Reference-guided *de novo* assembly	(Wang *et al.*, 2018)
O. sativa	3010	Reference-guided *de novo* assembly	(Sun *et al.*, 2016)
Brassica napus	53	Reference-based mapping and draft assembly of unmapped reads	(Hurgobin *et al.*, 2018)
B. oleracea	10	Reference-based mapping and draft assembly of unmapped reads	(Golicz *et al.*, 2016b)
B. rapa	3	*De novo* sequencing and assembly	(Lin *et al.*, 2014)
Triticum aestivum	18	Reference-based mapping and draft assembly of unmapped reads	(Montenegro *et al.*, 2017)
Zea mays	503	*De novo* transcriptome assembly	(Hirsch *et al.*, 2014)
Glycine soja	7	*De novo* draft genome assemblies	(Li *et al.*, 2014)
Medicago truncatula	15	*De novo* sequencing and assembly	(Zhou *et al.*, 2017)
Poplar (*Populus* spp.)	3	*De novo* sequencing and assembly	(Pinosio *et al.*, 2016)

They assembled 15.8 Mb (japonica) and 24.6 Mb (indica) of additional sequences, an increase of 4% and 6% in genome size, respectively (Yao *et al.*, 2015). Golicz *et al.* (2016b) constructed the pan-genome of *B. oleracea* from nine different morphotypes and assembled an additional 99 Mb of sequence. Similarly, in the *B. napus* pan-genome, 194 Mb additional sequences have been assembled (Hurgobin *et al.*, 2018). In the *Medicago truncatula* (a legume model plant) pan-genome, a total of 63 Mb non-redundant novel sequences were identified, with 47% (30 Mb) present in two or more accessions and 53% (33 Mb) being specific to a single accession (Zhou *et al.*, 2017). These additional sequences are the key for identification of novel and rare genomic structural variations within a plant species. For instance, analysing PAVs of genes related to agronomic traits within a pan-genome can help to identify genes that are uniquely present and absent in each individual. This information can then be used in evolutionary studies.

Golicz *et al.* (2016b) placed the number of present/absent genes on each branch of the *B. oleracea* pan-genome phylogenetic tree, assisting in the better visualization of uniquely present and absent genes, which can also represent the distance between the individuals. The extra information that pan-genomes add to the phylogenetic tree can help with the better understanding of the evolutionary patterns and selection among/within the species and can be linked to agronomic traits important for the breeding of advanced varieties.

Importance of Pan-genomes in Plant Breeding

Gene structural variations (PAVs and CNVs) related to agronomic traits have been frequently studied and reviewed within important crop species, including resistance genes in *Brassica* species (Xu *et al.*, 2012; Dolatabadian *et al.*, 2017; Bayer *et al.*, 2018b; Hurgobin *et al.*, 2018), yield and grain quality in rice and maize (Liu *et al.*, 2015; Wang *et al.*, 2015; Bai *et al.*, 2017) and flowering time in tomato and barley (Nitcher *et al.*, 2013; Würschum *et al.*, 2015). In addition, it has been reported that the dispensable regions of genomes are enriched with genes associated with agronomic traits, such as biotic stress and, especially, abiotic stress (McHale *et al.*, 2012; Li *et al.*, 2014; Golicz *et al.*, 2016b; Bayer *et al.*, 2018b). The pan-genome, by providing a platform for analysis and identification of SVs, facilitates the genetic dissection of the underlying genomic basis of agronomic traits. Knowledge of the gene content of individuals is important for genome-editing approaches as knowing what is present is a requirement to assess both on- and off-target edits (Scheben *et al.*, 2017). This will become increasingly important as genome editing-based breeding becomes more common (Scheben and Edwards, 2017, 2018). In the following sections, we highlight pan-genome applications in improving the accuracy and quality of SNPs calling and the identification of SVs within resistance genes, which are in high demand by breeders.

Detection of single nucleotide polymorphisms within a pan-genome

SNPs are widely used in plant breeding and genetic studies, including screening genetic diversity, marker-assisted selection (MAS), quantitative trait loci (QTL) mapping, genetic map construction and phylogenetic analysis. SNPs can be detected within or close to genes related to agronomic traits (Batley and Edwards, 2007).

Pan-genomes can assist with the identification of rare marker variation (e.g. SNPs) associated with QTL. The identification of SNPs across a genome is usually based on the alignment of sequence reads with a reference genome, and the identification of SNPs highly depends on the reference genome. Using the pan-genome, which reflects the gene content of the species rather than an individual, significantly increases the efficiency of SNP calling by including the regions displaying PAVs (Hurgobin and Edwards, 2017). This increases the chance of identification of SNPs linked to genes or QTL related to traits.

The identification of SNPs in a pan-genome can also be useful in phylogenetic studies for detecting accurate relationships between accessions without the ascertainment bias of using a reference from a single individual (Hurgobin and Edwards, 2017). Zhao *et al.* (2018) studied SNP variation within the rice pan-genome and found

evidence for introgressions from indica into tropical japonica. They reported an average of ~16.0% of the whole rice genome in tropical japonica may be an introgression from indica, where nine loci had a clear introgression pattern, including the thermotolerance allele of *OsTT1* (Os03g0387100) and the large-grain allele of *OsSPL13* (Os07g0505200) (Zhao *et al.*, 2018). These two loci have also been reported for introgression from indica to tropical japonica in other studies (Li *et al.*, 2015; Si *et al.*, 2016). Investigation of SNP variation within another rice pan-genome revealed thousands of genes had significantly lower diversity in cultivated rice when compared to the wild rice, which indicates candidate regions selected during domestication (Xu *et al.*, 2012). As increasing numbers of species-wide SNP studies are performed, there is a need to associate the resulting genotypic information with the pan-genome sequence within custom databases (Scheben *et al.*, 2019).

Understanding structural variations of resistance genes within a plant pan-genome

One of the main advantages of pan-genome construction is providing additional information regarding SVs, which are a major contributor to the evolution of the genetic diversity of genes related to agronomically important traits, such as resistance genes. In general, resistance genes refer to nucleotide-binding domain leucine-rich repeat (NLR) genes. In addition, receptor-like protein kinases (RLKs) and receptor-like proteins (RLPs) play roles in different plant mechanisms, including resistance responses (McHale *et al.*, 2006; Sekhwal *et al.*, 2015). In *Medicago* pan-genomes (consisting of 15 genotypes), different gene families show different patterns of CNVs, where NLRs show the higher number of CNVs in comparison to other gene families, such as zinc-finger motif proteins, RLKs and heat shock proteins (HSPs) (Zhou *et al.*, 2017). Similarly, in the poplar pan-genome, genes related to pathogen resistance were noticeably affected by CNVs (Pinosio *et al.*, 2016). In the *Glycine soja* pan-genome, genes associated with stress response, including NLRs and transcription factors, were considerably impacted by CNVs. In addition to resistance genes, CNVs among other types of genes related to agronomic traits, such as seed composition, flowering and maturity time, final biomass and organ size, have also been reported in the *G. soja* pan-genome (Li *et al.*, 2014). In the *Brassica rapa* pan-genome, Lin *et al.* (2014) reported copy number variations of genes related to the phenylpropanoid biosynthesis pathway (Lin *et al.*, 2014).

Golicz *et al.* (2016b) reported PAVs among resistance genes in the *B. oleracea* pan-genome, where from a total of 439 putative resistance genes, 251 were core genes and present in all genomes, and 188 of them were variable and missing from at least one genome. A similar pattern was observed in the *B. napus* pan-genome, where a total of 307 resistance genes have been identified, with 94 and 213 genes as core and variable, respectively (Hurgobin *et al.*, 2018). In *Brassica* species, resistance genes PAVs could be attributable to the large amounts of deletion and duplication that occurred during whole-genome duplication (WGD), whole-genome triplication (WGT) events and transposon-mediated gene duplication (Walker *et al.*, 1995; Franzke *et al.*, 2011; Lisch, 2013). Pan-genome studies can provide information to better understand the evolution of resistance genes within and between species and assist with the identification and mapping of novel resistance genes, which have been widely used in breeding programmes to improve resistance against devastating diseases. As with all genome analysis, accurate assembly and gene annotation are essential and this can be particularly problematic for the identification of disease resistance genes (Bayer *et al.*, 2017, 2018a).

Yao *et al.* (2015) reported PAVs of seven resistance/stress-related genes in the rice pan-genome, while Zhao *et al.* (2018) also reported that dispensable regions in the rice pan-genome (consisting of 66 genomes) were enriched with abiotic and biotic response genes, particularly for NLR (nucleotide-binding site–leucine-rich repeat) resistance gene. Xu *et al.* (2012) also reported CNVs in 14 disease resistance genes, where four contain a leucine-rich repeat domain and three contain a NB-ARC domain. In the wheat pan-genome, most of the variable genes were also associated with stress responses (Montenegro *et al.*, 2017). These observations indicate that many genes that are involved with biotic stress responses are variable.

Summary and Conclusions

Plant breeding approaches and strategies have been influenced substantially by the advances in genomic research. Plant pan-genomes can provide a complete genomic content of a species. As a single reference genome is insufficient to capture all genetic diversity present in a species, a plant pan-genome introduces a platform for comprehensive analysis of genetic diversity in a given species. Plant pan-genomes are particularly helpful for identification and characterization of structural genomic variations related to important agronomic traits and species-wide SNPs, both of which are of interest to breeders. In this chapter, we explore the progress achieved in plant pan-genomics and highlight its application in plant genetic and evolutionary studies. We also discuss the potential of pan-genomes for applications in plant breeding schemes.

Pan-genomes, which represent the genetic diversity within a species rather than a specific individual, add value to all aspects of genomic studies and molecular breeding strategies. Pan-genomics assists with genome evolutionary studies, comprehensive identification of genetic markers and the identification of novel agronomic trait-related genes and alleleic variants. Pan-genomes pave the way for the post-genomic era and will support the future of mining genomic variations and understanding of molecular mechanisms with direct applications for crop improvement.

References

Bai, X., Huang, Y., Hu, Y., Liu, H., Zhang, B., *et al*. (2017) Duplication of an upstream silencer of FZP increases grain yield in rice. *Nature Plants* 3, 885–893.

Batley, J. and Edwards, D. (2007) SNP applications in plants. In: Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., De Silva, H.N. (eds) *Association Mapping in Plants*. Springer, New York, pp. 95–102.

Bayer, P.E., Edwards, D. and Batley, J. (2018a) Bias in resistance gene prediction due to repeat masking. *Nature Plants* 4, 762–765.

Bayer, P.E., Golicz, A.A., Tirnaz, S., Chan, C.-K.K., Edwards, D., *et al*. (2018b) Variation in abundance of predicted resistance genes in the *Brassica oleracea* pangenome. *Plant Biotechnology Journal* 1–12.

Bayer, P.E., Hurgobin, B., Golicz, A.A., Chan, C.-K.K., Yuan, Y., *et al*. (2017) Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnology Journal* 15, 1602–1610.

Brunner, S., Fengler, K., Morgante, M., Tingey, S. and Rafalski, A. (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *The Plant Cell* 17, 343–360.

Dolatabadian, A., Patel, D.A., Edwards, D. and Batley, J. (2017) Copy number variation and disease resistance in plants. *Theoretical and Applied Genetics* 130, 2479–2490.

Franzke, A., Lysak, M.A., Al-Shehbaz, I. A., Koch, M.A. and Mummenhoff, K. (2011) Cabbage family affairs: The evolutionary history of Brassicaceae. *Trends In Plant Science* 16, 108–116.

Golicz, A.A., Batley, J. and Edwards, D. (2016a) Towards plant pangenomics. *Plant Biotechnology Journal* 14, 1099–1105.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., *et al*. (2016b) The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* 7, 13390.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., *et al.* (2014) Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell* 26, 121–135.

Hurgobin, B. and Edwards, D. (2017) SNP discovery using a pangenome: Has the single reference approach become obsolete? *Biology* 6, 21.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.-K.K., Tirnaz, S., *et al.* (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal* 16, 1265–1274.

Li, X.-M., Chao, D.-Y., Wu, Y., Huang, X., Chen, K., *et al*. (2015) Natural alleles of a proteasome A2 subunit gene contribute to thermotolerance and adaptation of african rice. *Nature Genetics* 47, 827–833.

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., *et al*. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* 32, 1045–1052.

Lin, K., Zhang, N., Severing, E.I., Nijveen, H., Cheng, F., *et al*. (2014) Beyond genomic variation-comparison and functional annotation of three *Brassica rapa* genomes: A turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics* 15, 250.

Lisch, D. (2013) How important are transposons for plant evolution? *Nature Reviews*. *Genetics* 14, 49–61.
Liu, L., Du, Y., Shen, X., Li, M., Sun, W*., et al.* (2015) Krn4 controls quantitative variation in maize kernel row number. *PLOS Genetics* 11, E1005670.
McHale, L., Tan, X., Koehl, P. and Michelmore, R. (2006) Plant NBS-LRR proteins: Adaptable guards. *Genome Biology* 7, 212.
McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E*., et al.* (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology* 159, 1295–1308.
Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., *et al*. (2013) Increased copy number at the Hvft1 locus is associated with accelerated flowering time in barley. *Molecular Genetics and Genomics* 288, 261–275.
Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., *et al*. (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Molecular Biology and Evolution* 33, 2706–2719.
Saxena, R.K., Edwards, D. and Varshney, R.K. (2014) Structural variations in plant genomes. *Briefings in Functional Genomics* 13, 296–307.
Scheben, A. and Edwards, D. (2017) Genome editors take on crops. *Science* 355, 1122–1123.
Scheben, A. and Edwards, D. (2018) Bottlenecks for genome-edited crops on the road from lab to farm. *Genome Biology* 19, 178.
Scheben, A., Wolter, F., Batley, J., Puchta, H. and Edwards, D. (2017) Towards CRISPR/Cas crops – bringing together genomics and genome editing. *New Phytologist* 216, 682–698.
Scheben, A., Verpaalen, B., Lawley, C.T., Chan, C.-K.K., Bayer, P.E*., et al.* (2019) CropSNPdb: A database of SNP array data for *Brassica* crops and hexaploid bread wheat. *The Plant Journal* 98, 142–152.
Sekhwal, M.K., Li, P., Lam, I., Wang, X., Cloutier, S., *et al*. (2015) Disease resistance gene analogs (RGAs) in plants. *International Journal of Molecular Sciences* 16, 19248–19290.
Si, L., Chen, J., Huang, X., Gong, H., Luo, J*., et al.* (2016) OsSPL13 controls grain size in cultivated rice. *Nature Genetics* 48, 447–456.
Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., *et al*. (2016) RPAN: Rice pan-genome browser for ~ 3000 rice genomes. *Nucleic Acids Research* 45, 597–605.
Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., *et al*. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences* 102, 13950–13955.
Walker, E.L., Robbins, T., Bureau, T., Kermicle, J. and Dellaporta, S. (1995) Transposon-mediated chromosomal rearrangements and gene duplications in the formation of the maize R-r complex. *The Embo Journal* 14, 2350–2363.
Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., *et al*. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49.
Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., *et al*. (2015) Copy number variation at the Gl7 locus contributes to grain size diversity in rice. *Nature Genetics* 47, 944–948.
Würschum, T., Boeven, P.H., Langer, S.M., Longin, C.F. H. and Leiser, W.L. (2015) Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genetics* 16, 96.
Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., *et al*. (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* 30, 105–111.
Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., *et al*. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology* 16, 187.
Yu, J., Wang, J., Lin, W., Li, S., Li, H., *et al*. (2005) The genomes of *Oryza sativa*: A history of duplications. *PLOS Biology* 3, E38.
Yu, J., Golicz, A.A., Lu, K., Dossa, K., Zhang, Y., *et al*. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal* 17, 881–892.
Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., *et al*. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics* 50, 278–284.
Zhou, P., Silverstein, K.A., Ramaraj, T., Guhlin, J., Denny, R., *et al*. (2017) Exploring structural variation and gene family architecture with de novo assemblies of 15 *Medicago* genomes. *BMC Genomics* 18, 261.

4 Genome Editing Technologies for Crop Improvement

Michael Pillay*
Vaal University of Technology, Vanderbijlpark, South Africa

Introduction

Progress in plant breeding from the domestication of the first crops to the present-day cultivars was made possible by using several major technologies, including cytogenetics, tissue culture, mutation breeding, transgenics and antisense RNA or RNA interference. Each of these technologies helped revolutionize crop breeding and enhance food production. However, some of these tools also came with their own list of shortcomings. For example, random mutagenesis produced many undesirable mutations and genome rearrangements; screening of germplasm with such tools is both expensive and laborious (Arora and Narula, 2017). Although transgenic technologies provide a powerful tool for crop improvement, public apprehension concerning potential food safety and gene flow has hampered the technology (Chen *et al.*, 2018). Other significant barriers in transgenic plant breeding are summarized by Zhong (2001). Transgenes may not follow Mendelian segregation; their expression can be affected by integration positions and structures of the transgenic DNA in host genomes. Transgenes may become unstable across generations, in different genetic backgrounds and environmental conditions, and could negatively affect gene expression (Zhong, 2001). Despite the promise of transgenic technology, only a few traits, such as herbicide and insect resistance, have reached the commercialization stage (Scheben *et al.*, 2017).

Across time, new genome editing techniques have become available. Initially, engineered nucleases were used as tools to target specific DNA sequences to edit genes precisely in model plants and crop plants (Osakabe and Osakabe, 2015). The engineered nucleases induced double-stranded DNA breaks at the target site that are then repaired by natural processes of homologous recombination (HR) or non-homologous end joining (NHEJ). Four major types of genome editing technologies are currently available: (i) engineered homing endonucleases/meganucleases; (ii) zinc finger nucleases (ZFNs); (iii) transcription activator-like effector nucleases (TALENs); and (iv) clustered regularly interspersed short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) (CRISPR/Cas9) (Osakabe and Osakabe, 2015). Meganucleases are unique enzymes with high activity and long recognition sequences (>14 bp) and are able to produce site-specific digestion of target DNA (Smith *et al.*, 2006). One of the major problems of meganucleases is the need for the introduction of a long cleavage site in the region of interest (Manesh and Malik, 2016). ZNFs are chimeric proteins

* Email: mpillay@vut.ac.za

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

composed of a synthetic zinc-finger-based DNA-binding domain and DNA cleavage domain of the endonuclease *Fok*1. ZNFs can be designed to cleave almost any long stretch of double-stranded DNA by modification of the zinc finger DNA-binding domain (Durai *et al.*, 2005). There are several disadvantages of using ZNFs, including the fact that it is time consuming, it is costly to construct the target enzymes, and it has low specificity and high off-target mutations (Ma *et al.*, 2016). TALENs turned out to be a substitute for ZNFs and were identified as restriction enzymes that could be manipulated for cutting specific DNA sequences (Arora and Narula, 2017). Traditionally, TALENs were considered to be long segments of activator-like effectors (TALEs) that occur naturally and are joined to the domain of the enzyme *Fok*1 (Jaganathan *et al.*, 2018). Although they are easier to use than ZNFs, TALENs require construction of complicated tandem repeat domains in the transcription activator-like (TAL) proteins and, since TALEs contain a large number of repeat domains, it can be difficult to synthesize new variants (Ma *et al.*, 2016).

Recently, a newly established editing tool, CRISPR/Cas9, has superseded previous techniques and has been utilized widely for editing the genomes of various organisms, including bacteria, yeast, animals and plants (Osakabe and Osakabe 2015). CRISPR can be divided into three major types, I, II and III, and several subtypes (Arora and Narula, 2017) with different Cas genes. The type II system requires only one Cas protein to recognize and cleave DNA that matches a single guide RNA (sgRNA), whereas the others require a set of proteins. CRISPR/Cas9 technology involves three main components: the Cas nuclease, single guide RNA (sgRNA) that is composed of a CRISPR RNA (crRNA) and a transactivating crRNA (tracRNA), and target sites upstream of the protospacer adjacent motif (PAM; aNy base-Guanosine-Guanosine (NGG)). The Cas9 nuclease associated with CRISPR makes double-stranded cuts in the target DNA. The palindromic sequences in CRISPR are 29 nucleotides repeat sequences separated by various 32-nucleotide spacer sequences that were first reported in bacteria (Song *et al.*, 2016). The CRISPR/Cas9 cleavage methodology requires: (i) a short synthetic guide RNA (sgRNA) sequence of about 20 nucleotides that binds to the target DNA; and (ii) Cas9 enzyme that cleaves 3–4 bases after the protospacer adjacent motif (PAM, which is usually 5′NGG) (Li *et al.*, 2016). Implementing the CRISPR/Cas9 technology requires: (i) identifying the PAM sequence in the target gene; (ii) synthesizing a single gRNA; (iii) cloning the gRNA into a suitable vector; (iv) introducing the vector into suitable cells; (v) screening; and (vi) validating the edited lines (Jaganathan *et al.*, 2018).

CRISPR/Cas9 is considered to be simple, economical and versatile in many applications (Song *et al.*, 2016). It has received extensive attention because of its easy manipulation, high efficiency and wide application in gene mutation and transcriptional regulation in plants (Liu *et al.*, 2017).

The first reports of CRISPR/Cas9 editing in plants appeared in 2013 (Brooks *et al.*, 2014). Early reports on the CRISPR/Cas9 technology demonstrated the efficacy of the system in the model plants, *Arabidopsis thaliana* and *Nicotiana benthamiana* (Brooks *et al.*, 2014). The first three crop species in which CRISPR/Cas9 technology was demonstrated included rice (Zhang *et al.*, 2014), sorghum (Jiang *et al.*, 2013) and wheat (Wang *et al.*, 2014). Since then, there has been an explosion in research in genome editing of plants using the CRISPR/Cas9 technology and its modifications. It is not possible to review all the research on the value of the CRISPR/Cas9 technology for plant improvement, and in this chapter, I have selected some of the major themes in which the technique has been used in plants. Several excellent reviews have highlighted the value of the CRISPR/Cas9 system in different aspects of crop improvement. A few of these reviews (Ceasar *et al.*, 2016; Ma *et al.*, 2016; Song *et al.*, 2016; Arora and Narula, 2017; Liu *et al.*, 2017; Scheben *et al.*, 2017) also provide excellent figures depicting the mechanisms of the CRISPR/Cas9 technology. Therefore, I will not consider the mechanisms involved in the CRISPR/Cas9 technology but will concentrate on some of the ways in which this technology has enhanced crop improvement.

The human population is predicted to be close to 10 billion by 2050, which necessitates that global food production increase by 60–100% (FAO, 2016, cited in Jaganathan *et al.*, 2018). At the current rate of improvement, yields of the staple crops, viz., maize, rice, wheat and soybean, are expected to increase by about 38–67% (Ray *et al.*, 2013). Crops are also susceptible to a large number of pathogens, including viruses, bacteria and fungi, which are responsible for

important economic losses that range from 20–40% of global agricultural production (Savary *et al*., 2012). In addition, the impact of climate change on agriculture requires crops with greater yields and tolerance to abiotic stresses (Scheben *et al*., 2017). In addition to traditional crop improvement, new technologies are required to enhance crop improvement. CRISPR/Cas9 holds promise for contribution to crop improvement by direct genome editing. Current research using the CRISPR/Cas9 technology has already shown that disease resistance can be modified by targeting plant pathogens. In addition, CRISPR/Cas9 has been used in many other areas in crop improvement, including yield, nutrition, domestication and other aspects, some of which are discussed in this chapter.

Plant Disease Resistance

Generally, plant diseases have been controlled by the development of resistant varieties and by using agrochemicals (Borrelli *et al*., 2018). However, it is well known that plant pathogens are constantly evolving and resistance to diseases breaks down even though plants have evolved sophisticated mechanisms to resist plant pathogens (Wang *et al*., 2016). In addition, agrochemicals can cause environmental contamination and are a health hazard to humans. Traditional breeding programmes for disease resistance involve the incorporation of resistance genes from wild species into elite cultivars. This method often transfers unnecessarily large genomic regions into the elite cultivar, changing its characteristics. For example, this is a concern in banana breeding when elite cultivars are crossed within the diploid wild species, *Musa acuminata,* that harbour genes for disease and pest resistance (Pillay and Tripathi, 2007). Several crosses and backcrosses are required to reach the desired end product, extending the time required to combine disease resistance with acceptable agronomic traits. The CRISPR/Cas9 system has been employed to overcome several agricultural challenges, including biotic stress resistance. While resistance to viruses has been most successful, genome editing tools have also been used to improve fungal and bacterial disease resistance (Borrelli *et al*., 2018). For example, CRISPR/Cas9 was used to mutate one of the three *MLO* alleles in bread wheat. Thereafter, the species showed improved resistance to powdery mildew (*Blumaria graminis* f. sp. *tritici*) infection. Targeting the same genes in tomato plants created complete resistance to *Oidium neolycopersici*, the causative agent of powdery mildew in tomato (Borrelli *et al*., 2018). Plants resistant to rice blast disease were generated through CRISPR/Cas9 by disrupting the *OsERF922* and *OsSEC3A* genes (Wang *et al*., 2016; Ma *et al*., 2018). One of the advantages of the CRISPR/Cas9 system is that it can be used for the enhancement of disease resistance even in perennial crops. Several examples of the application of the CRISPR/Cas9 in perennial plants are reported in Borrelli *et al*. (2018).

CRISPR/Cas9 for resistance to plant viruses

Plant viruses infect many crops from cereals to vegetables, limiting crop yield and posing a serious threat to food security (Khatodia *et al*., 2017). CRISPR/Cas9 has been used for the development of virus-resistant plants. Most studies involving the CRISPR/Cas9 system in plants have targeted virus resistance. The CRISPR/Cas9 system has been used successfully to create resistance against both plant DNA and RNA viruses. The majority of the plant virus studies using CRISPR/Cas9 technology have targeted the single-stranded DNA geminivirus genomes. Geminiviruses in the genus *Begomovirus* are a large family of plant DNA viruses that cause severe crop losses and threaten food security worldwide (Ali *et al*., 2015). Resistance to geminiviruses using CRISPR/Cas9 was initially studied in the model plants *N. benthamiana* and *Arabidopsis* (Ji *et al*., 2015; Baltes *et al*., 2015). Ji *et al*. (2015) demonstrated resistance against beet severe curly top virus (BSCTV) by creating mutations in the viral target sequences. Baltes *et al*. (2015) produced resistance against bean yellow dwarf virus (BeYDV). Ali *et al*. (2015) established the efficacy of the CRISPR/Cas9 system targeting the tomato yellow leaf curl virus (TYLCV) and BSCTV in the model plant *N. benthamiana*. Mahas and Mahfouz (2018) also engineered resistance to TYLCV (a devastating virus affecting a number of crops, including tomato, *Solanum lycopersicum*) using CRISPR/Cas9.

The classical DNA targeting CRISPR/Cas9 system cannot be used to target RNA viruses because the Cas9 from *Streptococcus pyogenes* can only recognize double-stranded DNA (Borrelli *et al.*, 2018). Plant RNA viruses require host factors to maintain their life cycle (Chandrasekaran *et al.*, 2016). Modifications of the CRISPR/Cas9 system have been used to engineer interference against RNA viruses in plants. The C2c2 and FnCas9 CRISPR/Cas systems were used to develop resistance to RNA viruses in plants (Green and Hu, 2017). CRISPR/Cas13a was shown to be effective against turnip mosaic virus (TuMV) (Aman *et al.*, 2018). CRISPR/Cas9 has been effectively used in creating virus resistance in cucumber (Chandrasekaran *et al.*, 2016). Immunity was exhibited against cucumber vein yellowing virus, zucchini yellow mosaic virus and papaya ring spot mosaic virus. Of interest is that many of the CRISPR/Cas9 systems have been used to demonstrate virus resistance in the model plants *Arabidopsis* and *N. benthamiana*. It would be interesting to see if this technique will be effective to create virus resistance in specific plants.

CRISPR/Cas9 for developing bacterial resistance

Genome editing has been recognized as one of the main tools in developing disease resistance in plants (Borrelli *et al.*, 2018) as resistance is usually achieved by modification of a single gene. There are a few published reports on the use of CRISPR/Cas9 to counteract crop bacterial diseases (Borrelli *et al.*, 2018). In rice, CRISPR/Cas9 mutagenesis of the ethylene-responsive factor (ERF) transcription factor gene *OsERF922* enhanced rice blast resistance (Wang *et al.*, 2016). Jia *et al.* (2016) achieved resistance to citrus bacterial canker (CBC) caused by *Xanthomanas citri* subsp. *citri* in Duncan grapefruit by editing the effector-binding element in the promoter of the *Lateral Organ Boundaries 1* gene. Resistance to the same disease was also enhanced in Wanjinchen oranges by deleting a sequence from both the alleles ($CsLOB1^{G}$ and $CsLOB1^{-}$) (Peng *et al.*, 2017).

The long-term success of CRISPR/Cas9 technology in plant protection is dependent on new scientific knowledge. The technology can only be used if one knows which genes to modify and which modification to carry out to create disease-resistant plants (Borrelli *et al.*, 2018).

The optimism for the use of the CRISPR/Cas9 technology in disease resistance is summarized in Borrelli *et al.* (2018). The scientific knowledge of pathosystems is now advanced to such a level that specific genes can be edited. Disease resistance can be achieved by editing a single gene. The mutations created by CRISPR/Cas9 technology are now well understood to create disease resistance in plants.

Genome Modification for Nutritional Improvement

The demand for more nutritious crops is on the rise as the incomes of households increase. Genome editing techniques have made it possible to enhance the food quality and value of crops by increasing their nutritional status. Lycopene is considered to be a crucial component in treating chronic diseases and lowering the risk of cancer and cardiovascular diseases (Li *et al.*, 2018). One of the most commonly eaten fruits worldwide is the tomato. The amount of lycopene in tomato is an important quality trait with industrial, health and nutritional attributes (Li *et al.*, 2018). Increasing the lycopene content of tomato has been of interest to make the fruit more nutritious. Li *et al.* (2018) were able to increase the lycopene content of tomato by 5.1-fold via the CRISPR/Cas9 system using *Agrobacterium*-mediated transformation. This achievement was possible even though complex pathways are involved in lycopene formation in plants.

Jiang *et al.* (2017) used CRISPR/Cas9 technology to increase the oleic acid content in the seeds of *Camelina sativa* from 16% to >50% of the fatty acid composition. The advantage of this increase was that it significantly decreased the content of linoleic and linolenic acids that are considered undesirable polyunsaturated fatty acids. With these changes, the oil from the seeds of *C. sativa* has become a healthier product, is more oxidatively stable and has become better suited to the production of biofuels. High-amylose rice was created through CRISPR/Cas9 that targeted the starch branching enzyme that is coded by the genes *SBEl* and *SBEllb* (Sun *et al.*, 2017). This rice is high in resistant starch and is beneficial to

human health, especially to people with diabetes. In addition to having a low glycaemic index, resistant starch products are not digested and absorbed in the stomach or small intestine and are passed directly to the large intestine. Although wheat products are one of the major components of the human diet, the gluten products in wheat trigger certain pathologies in susceptible individuals. Sánchez-León *et al.* (2018) developed a low-gluten wheat with CRISPR/Cas9 by targeting the α-gliadin gene family. It is envisaged that this product will have reduced immunoreactivity for gluten-intolerant consumers.

Plant foods contain most of the mineral nutrients required for the human body (Grusak and DellaPenna, 1999), but these are often not present in sufficient amounts (Vasconcelos *et al.*, 2003). Micronutrient malnutrition, particularly vitamin A, Fe and Zn deficiencies, affects billions of people in the world (Wang *et al.*, 2011). Iron deficiency affects about 2 billion people and iron anaemia is responsible for a fifth of childhood deaths and a tenth of maternal mortalities (Black, 2003). Similarly, health problems attributable to zinc deficiency include anorexia, dwarfism, weak immune system, skin lesions, hypogonadism and diarrhoea (McClain *et al.*, 1992). Improving the micronutrient composition of plant foods may become a sustainable strategy to combat deficiencies in human populations, replacing or complementing other strategies, such as food fortification or nutrient supplementation (Hess, 2013). The iron and zinc content of plants could be increased by biofortification, which is a strategy to increase the nutrient content of staple foods through agricultural means, including breeding, genetic engineering, mutagenesis and agronomic approaches (Hotz, 2013). With the success obtained in increasing the lycopene content in tomato, it is envisaged that the CRISPR/Cas9 system may play a major role in improving the micronutrient content of plants.

Yield of Crop Plants

Crop growth and yield are controlled by several phytohormones (Miao *et al.*, 2018). One of the key hormones is abscisic acid (ABA), which controls plant growth and stress responses, such as drought, salinity, osmotic stress, extreme temperature, pathogen attack, and so on (Cutler *et al.*, 2010; Lee and Luan, 2012). ABA binds to the ABA receptor pyrobactin resistance 1/PYR1-like/regulatory components of the ABA receptor (PYR1/PYL/RCAR) proteins, which are usually referred to PYLs (Zhang *et al.*, 2017). The CRISPR/Cas9 technology was used to edit two groups of the PYL genes in rice (Miao *et al.*, 2018). One of the groups exhibited better growth and improved grain productivity by up to 31% compared with traditional breeding methods. This study showed that it was possible to use the CRISPR/Cas9 technology to create mutations in a subfamily of the ABA receptor genes to enhance rice grain yield. Similar research may have the potential to improve yields in other crops.

In view of the impending climate change, developing drought-tolerant crops is necessary to meet the demand for food, feed and fuel (Shi *et al.*, 2017). The phytohormone ethylene plays an important role in regulating the responses of plants to abiotic stresses. In maize, the *ARGOS8* gene is a negative regulator of ethylene biosynthesis (Shi *et al.*, 2017). Overexpression of the *ARGOS8* gene in transgenic maize increased grain yield under drought conditions (Shi *et al.*, 2015). The CRISPR/Cas 9 technology was used to develop several variants of the *ARGOS8* gene by changing the DNA sequence of the gene in maize. These plants were tested under stress conditions at flowering and the results showed that maize yields increased by five bushels per acre (one bushel = 56 pounds = 25.40 kg). The plants showed no yield loss when grown under optimal, well-watered conditions. This study showed that CRISPR/Cas9 genome editing of a single gene was able to create novel variants in maize that had a positive effect on a complex trait, such as drought tolerance (Shi *et al.*, 2017).

Plant Domestication

The domestication of new crops can promote agricultural diversity and provide a solution to many of the problems associated with intensive agriculture (Osterberg *et al.*, 2017). Only about 200 plant species out of more than 300,000 are commercially important. The majority of nutrients consumed by humans come from just three crops – rice, wheat and maize. There are many neglected or orphan crops, semi-domesticated and wild plants that could become potential food

sources for humans. Some of these plants, such as Bambara groundnut, amaranth and quinoa, can grow in environments that are not conducive to rice, wheat and maize cultivation. Bambara groundnut can be cultivated under drought conditions. As a legume, it is also able to fix nitrogen and grow in low-input and low-nutrient soils (Unigwe *et al.*, 2017). Some species in the genus *Bromus*, which is closely related to our cereals, have not been exploited for human use (Pillay, 1995). The Tarahumara Indians in northern Mexico used the grains of some native *Bromus* species to aid fermentation in making one of their cultural beverages (https://en.wikipedia.org/wiki/Bromus). One of the species, *Bromus catharticus*, having very large seeds, could be domesticated to increase seed size of other species (personal observation). With genetic improvement, the yield of *B. catharticus* is envisaged to be comparable to that of wheat and barley.

Genes have been identified for traits associated with domestication (Osterberg *et al.*, 2017). These are single genes that have a marked effect on the domestication-associated phenotype. Sequencing of the genomes of model plants, crops and underutilized crop and wild plants is important for using genome editing techniques to change the phenotype (Schreiber *et al.*, 2018). Knowledge of the genome sequence of these plants would make it possible to use genome editing techniques such as CRISPR/Cas9 to accelerate domestication. Genome editing has already been used to exploit the genetic diversity in wild plants and transform some wild species into potential cash crops.

Ground cherries (*Physalis pruinosa*) and their close relative, the Cape or golden goose berries (*Physalis peruviana*), are grown in many parts of the world. Some of their traits, such as preharvest fruit drop, have made them unattractive for large-scale production (Lemmon *et al.*, 2018). It has been shown that by editing two genes using the CRISPR-Cas9 in ground cherries, the plants not only produced larger fruits but also increased the yield because the fruits were about 24% heavier.

In another study, Zsögön *et al.* (2018) reported that editing of six loci that are important for yield and productivity in the present-day tomato crop lines enabled *de novo* domestication of wild *Solanum pimpinellifolium* using CRISPR/Cas9. *Solanum pimpinellifolium*, because of its close relationship to *S. lycopersicum*, has been a genetic source for many commercially important tomato traits. It is a wild species found in the coastal areas of Peru and Ecuador (Zuriaga *et al.*, 2009). Genome editing resulted in an altered morphology of *S. pimpinellifolium* and an increase in the size, number and nutritional value of the fruits. The fruit lycopene accumulation improved by 500%.

Medicinal Plants

Research on the identification of useful compounds from plants has soared during the past few years. Antibiotic-resistant microorganisms are on the increase perhaps because of the injudicious use of these drugs. There is a greater awareness of the need to seek new effective plants, plant constituents or antimicrobial agents to treat diseases caused by pathogenic microorganisms. Plants contain bioactive, non-nutrient and biologically active compounds, such as terpenoids, saponins, flavonoids, coumarins, tannins, phenols and cardiac glycosides. The health benefits of some of these phytocompounds are discussed by Madike *et al.* (2017). For example, flavonoids have been reported to possess a wide variety of biological activities, which include antimicrobial, anti-inflammatory, anti-angionic, analgesic, anti-allergic, cytostatic and antioxidant, antiviral, anti-carcinogenic, anticancer and antidiarrhoeal properties. They also act as free-radical scavengers, which prevent oxidative cell damage and diseases associated with oxidative damage to membranes, proteins and DNA (Madike *et al.*, 2017). The genomes of some medicinal plants have been completely sequenced (Liu *et al.*, 2017). It is envisaged that the CRISPR/Cas9 system could be used to edit targeted genes in medicinal plants, study the synthesis of effective compounds, select traits for increased yield, and promote research on biosynthetic pathways and regulatory mechanisms (Liu *et al.*, 2017).

Some Advantages of CRISPR/Cas9 for Genome Editing in Plants

Compared with earlier techniques, the CRISPR/Cas9 technology for genome editing is considered simple, flexible, versatile and efficient for plant

improvement (Ma *et al.*, 2016). The technology is expected to have a great impact on basic and applied research in plant biology (Song *et al.*, 2016). The precise modification of genes in elite cultivars will save the time-consuming task of backcrossing that is usually involved in conventional breeding schemes. Another advantage of CRISPR/Cas9 technology is that it can edit multiple target genes simultaneously with a single molecular construct (Liu *et al.*, 2017; Borrelli *et al.*, 2018). For example, a single construct was used to create mutations in 14 genes in *Arabidopsis* (Borrelli *et al.*, 2018). Highly efficient multisite genome editing was achieved in allotetraploid cotton using CRISPR/Cas9 (Wang *et al.*, 2018) and efficient multiallelic mutagenesis was achieved in potato (Andersson *et al.*, 2017). Simultaneous editing of multiple genomes in plants has many applications, such as studying multiple related genes or knockout of functionally redundant genes, or genetic improvement of multiple traits in crop breeding (Ma *et al.*, 2016). This is a major advantage of this gene editing technology because it is known that the vast majority of traits in plants are polygenic and are, as such, controlled by the action of several genes.

The CRISPR/Cas9 system is a non-transgenic technology and is considered to be applied to asexually propagated, heterozygous perennial plants (Chen *et al.*, 2018). It is still early to predict the regulations that will apply to CRISPR/Cas9 technology and whether the plants altered with this technology will be considered genetically modified organisms (GMOs).

Some Disadvantages of CRISPR/Cas9 for Plant Genetic Improvement

Although regarded as a powerful technique for gene manipulation, some of the weakness of CRISPR/Cas9 technology are discussed in Song *et al.* (2016) and Peng *et al.* (2016). Similar to ZNFs and TALENs, CRISPR/Cas9 has the problem of off-target effects that may introduce unexpected mutations. These off-targets appear to be affected by the concentration ratio between the enzyme Cas9 and sgRNA. The higher the Cas9 : sgRNA ratio, the more severe the effect of pairing of the sgRNA to non-specific sequences in the genome (Hsu *et al.*, 2013; Pattanayak *et al.*, 2013; Peng *et al.*, 2016). Optimal mutagenesis was obtained in *Arabidopsis* when the ratio of Cas9 to some of the targeted genes was 1:1. Promiscuous PAM sites could lead to undesired cleavage of DNA regions (Sternberg *et al.*, 2014). Insufficient Cas9 codon optimization may lead to inefficient translation of Cas9 proteins in target species (Hsu *et al.*, 2013).

Although some researchers have considered CRISPR/Cas9 technology to be a simple technique, sgRNA design is a major concern in its application (Peng *et al.*, 2016). Initially, it was assumed that Cas9/rgRNA complexes would cleave double-stranded DNA in the presence of PAM and a complementary target sequence. It is now known that some sgRNAs are less efficient and even inactive, making the screening and testing of sgRNAs a requirement in CRISPR/Cas9 research.

One of the limitations in the application of the CRISPR/Cas9 system is the ability to deliver the genome editing tools into plant cells and the regeneration of plants from such cells (Altpeter *et al.*, 2017). Many plants are recalcitrant to tissue-culture techniques and will not be amenable to being modified by the CRISPR/Cas9 system of genome editing. Efficient systems to deliver genome editing tools into plant cells must be developed (Osakabe and Osakabe, 2015). To overcome this problem, viral vectors have been used to deliver sgRNA of the CRISPR/Cas9 system in transgenic *N. benthamiana* (Green and Hu, 2017). The CRISPR/Cas9 technology can only be used in plants that have their genomes or a large part of their genomes sequenced or have known gene sequences to develop sgRNAs that would target specific genes.

A recent study in human biology found that many of the cells modified with the CRISPR/Cas9 technology had large genetic rearrangements, such as deletions and insertions, that may lead to switching on or off of genes (Kosicki *et al.*, 2018). Further research is needed to investigate the extent of off-target mutations and cleavage efficiency of the CRISPR/Cas9 system in plants (Bortesi *et al.*, 2015).

Conclusion

The potential of CRISPR/Cas9 for genome editing is not yet fully exploited. Modifications to increase the efficiency, specificity and range of accessible

targets of the CRISPR/Cas9 are on the rise (Ma *et al*., 2016). Orthologues of *spCAS9*, such as *stCas9* and *saCas9*, have been confirmed to be functional in plants. The newly developed *Cpf1* can create double-stranded breaks with staggered ends (Zetsche *et al*., 2015). It appears that *Cpf1* may be better suited for genome editing than Cas9. One of the shortcomings of Cas9 is the blunt double-stranded cleavage and G-rich PAM requirement. Other advantages of *Cpf1* are addressed in Zetsche *et al*. (2015).

Recently, two research groups, one from Stanford University and the other from the Joint Initiative of Metrology and Biology, described a new genome editing tool known as MAGESTIC (multiplexed accurate genome editing with short, trackable, integrated cellular barcodes). MAGESTIC uses the CRISPR/Cas9 genome editing technology but adds two key elements: active recruitment of 'donor' DNA to the cleaved site and a genomic barcode (Roy *et al*., 2018).

MAGESTIC provides the cell with specific 'donor' DNA that the cell's DNA repair system can use as a template to replace the original sequence at the cleavage site (Roy *et al*., 2018). During DNA repair, a cell searches through millions of base pairs of DNA sequences to find the correct 'donor' DNA. The major advance with MAGESTIC is that it provides the cell with the 'donor' directly, making the system more precise and less error prone than CRISPR/Cas9 to edit genomes. CRISPR/Cas9 and its variant technologies will be useful in addressing diverse agricultural problems, such as increased yields and nutritional value, resistance to diseases and pests, overcoming abiotic stresses and improving other important agronomic traits (Jaganathan *et al*., 2018). A new revolution in agriculture is on our doorstep.

References

Ali, Z., Abulfaraj, A., Idris, A., Ali, S., Tashkandi, M., *et al*. (2015) CRISPR/Cas9-mediated viral interference in plants. *Genome Biology* 16, 238.

Altpeter, F., Kannan, B., Jung, J.H., Oz, T.M., Karan, R., *et al*. (2017) Genetic improvement of sugarcane by targeted loss or gain of function mutations using TALEN or CRISPR-Cas9. *Proceedings of the Plant and Animal Genome Conference (PAG XXV)*. San Diego, California, USA, 14–18 January 2017. W699.

Aman, R., Ali, Z., Butt, H., Mahas, A, Aljedaani, F., *et al*. (2018) RNA virus interference via CRISPR/Cas13a system in plants *Genome Biology* 19, 1.

Andersson, M., Turesson, H., Nicolia, A., Falt, A.S., Samuelsson, M., *et al*. (2017) Efficient targeted multiallelic mutagenesis in tetraploid potato (*Solanum tuberosum)* by transient CRISPR-Cas9 expression in protoplasts. *Plant Cell Reports* 36, 117–128.

Arora, L. and Narula, A. (2017) Gene editing and crop improvement using CRISPR-CAS 9 system. *Frontiers in Plant Science* 8, 1–21.

Baltes, N.J., Hummel, A.W., Konecna, E., Cegan, R., Bruns, A.N., *et al*. (2015) Conferring resistance to geminiviruses with the CRISPR–Cas prokaryotic immune system. *Nature Plants* 1, 145.

Black, R. (2003) Micronutrient deficiency – an underlying cause of morbidity and mortality. *Bulletin of the World Health Organization* 81, 79.

Borrelli, V.M.G., Brambilla, V., Rogowsky, P., Marocco, A. and Lanubile A. (2018) The enhancement of plant disease resistance using CRISPR/Cas9 technology. *Frontiers in Plant Science* 9, 1245.

Bortesi, L. and Fischer, R. (2015) The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances* 33, 41–52.

Brooks, C., Nekrasov, V., Lippman, Z.B. and Van Eck, J. (2014) Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. *Plant Physiology* 166, 1292–1296.

Ceasar, S.A., Rajan, V., Prykhozhij, S.V., Berman, J.N. and Ignacimuthu, S. (2016) Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. *Biochimica et Biophysica Acta* 1863, 2333–2344.

Chandrasekaran, J., Brumin, M., Wolf, D., Leibman, D., Klap, C., *et al*. (2016) Development of broad virus resistance in non-transgenic cucumber using CRISPR/Cas9 technology. *Molecular Plant Pathology* 17, 1140–1153.

Chen, L., Li, W., Katin-Grazzini, L., Ding, J., Gu, X., *et al*. (2018) A method for the production and expedient screening of CRISPR/Cas9-mediated non-transgenic mutant plants. *Horticulture Research* 5, 13.

Cutler, S.R., Rodriguez, P.L., Finkelstein, R.R. and Abrams, S.R. (2010) Abscisic acid: Emergence of a core signaling network. *Annual Review of Plant Biology* 61, 651–679.

Durai, S., Mani, M., Kandavelou, K., Wu, J., Porteus, M.H., *et al*. (2005) Zinc finger nucleases: Custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Research* 33, 5978–5990.

Green, J.C. and Hu, J.S. (2017) Editing plants for virus resistance using CRISPR-Cas. *Acta Virologica* 16, 138–142.

Grusak, M.A. and DellaPenna, D. (1999) Improving the nutrient composition of plants to enhance human nutrition and health. *Annual Review of Plant Physiology and Plant Molecular Biology* 50, 133–161.

Hess, S.Y. (2013) Zinc: Deficiency disorders and prevention programs. *Encyclopedia of Human Nutrition* 3rd edn. Academic Press, New York, pp. 431–436.

Hotz, C. (2013) Biofortification. *Encyclopedia of Human Nutrition*, 3rd edn. Academic Press, New York, pp. 175–181.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., *et al*. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* 31, 827–832.

Jaganathan, D., Ramasamy, K., Sellamuthu, G., Jayabalan, S. and Venkataraman, G. (2018) CRISPR for crop improvement: An update review. *Frontiers in Plant Science* 9, 985.

Ji, X., Zhang, H., Zhang, Y., Wang, Y. and Gao, C. (2015) Establishing a CRISPR–Cas-like immune system conferring DNA virus resistance in plants. *Nature Plants* 1, 15144.

Jia, H., Zhang, Y., Orbović, V., Xu, J., White, F.F., *et al*. (2016) Genome editing of the disease susceptibility gene *CsLOB1* in citrus confers resistance to citrus canker. *Plant Biotechnology Journal* 15, 817–823.

Jiang, W., Zhou, H., Bi, H., Fromm, M., Yang, B. *et al*. (2013) Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Research* 41, e188.

Jiang, W.Z., Henry, I.M., Lynagh, P.G., Comai, L, Cahoon, E.B., *et al*. (2017) Significant enhancement of fatty acid composition in seeds of the allohexaploid, *Camelina sativa*, using CRISPR/Cas9 gene editing. *Plant Biotechnology Journal* 15(5), 648–657.

Khatodia, S., Bhatotia, K. and Tuteja, N. (2017) Development of CRISPR/Cas9 mediated virus resistance in agriculturally important crops. *Bioengineered* 8, 274–279.

Kosicki, M., Tomberg, K. and Bradley, A. (2018) Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 36, 765–771.

Lee, S.C. and Luan, S. (2012) ABA signal transduction at the crossroad of biotic and abiotic stress responses. *Plant Cell Environment* 35, 5360.

Lemmon, Z.H., Reem, N.T., Dalrymple, J., Soyk, S., Swartwood, K.E., *et al*. (2018) Rapid improvement of domestication traits in an orphan crop by genome editing. *Nature Plants* 4, 766–770.

Li, J., Sun, Y., Du, J., Zhao, Y. and Xia, L. (2016) Generation of targeted point mutations in rice by a modified CRISPR/Cas 9 system. *Molecular Plant* 10, 526–529.

Li, X., Wang, Y., Chen, S., Tian, H., Fu, D., *et al*. (2018) Lycopene is enriched in tomato fruit by CRISPR/Cas9-mediated multiplex genome editing. *Frontiers in Plant Science* 9, 559.

Liu, X., Wu, S., Xu J., Sui, C. and Wei, J. (2017) Application of CRISPR/Cas 9 in plant biology. *Acta Pharmaceutica Sinica B* 7, 292–302.

Ma, J., Chen, J., Wang, M., Ren, Y., Wang, S., *et al*. (2018) Disruption of OsSEC3A increases the content of salicylic acid and induces plant defense responses in rice. *Journal of Experimental Botany* 69, 1051–1064.

Ma, X., Zhu, Q., Chen, Y. and Liu, Y-G. (2016) CRISPR/Cas9 platforms for genome editing in plants: Developments and applications. *Molecular Plant* 9, 961–974.

Madike, L.N., Takaidza, S. and Pillay, M. (2017) Preliminary phytochemical screening of crude extracts from the leaves, stems and roots of *Tulbaghia violacea*. *International Journal of Pharmacognosy and Phytochemical Research* 9, 1300–1308.

Mahas A. and Mahfouz, M. (2018) Engineering virus resistance via CRISPR–Cas systems. *Current Opinion in Virology* 32, 1–8.

Manesh, D.M. and Malik, P. (2016) The current state-of-the-art in therapeutic genome editing and the future. *Gene Technology* 5, 135.

McClain, C.J., Stuart, M.A., Vivian, B., McClain, M., Talwalker, R., *et al*. (1992) Zinc status before and after zinc supplementation of eating disorder patients. *Journal of the American College of Nutrition* 6, 694–700.

Miao, C., Xiao, L., Hua, K., Zou, C., Zhao, Y., *et al*. (2018) Mutations in a subfamily of abscisic acid receptor genes promote rice growth and productivity. *Proceeding of the National Academy of Sciences* 23, 6058–6063.

Osakabe, Y. and Osakabe, K. (2015) Genome editing with engineered nucleases in plants. *Plant Cell Physiology* 56(3), 389–400.

Osterberg, J.T., Xiang, W., Olsen, L.I., Edenbrandt, A.K., Vedel, S.E., *et al*. (2017) Accelerating the domestication of new crops: Feasibility and approaches. *Trends in Plant Science* 22, 373–384.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., *et al*. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA programmed Cas9 nuclease specificity. *Nature Biotechnology* 31, 839–843.

Peng, A., Chen, S., Lei, T., Xu, L., He, Y., *et al*. (2017) Engineering canker-resistant plants through CRISPR/Cas9-targeted editing of the susceptibility gene *CsLOB1* promoter in citrus. *Plant Biotechnology Journal* 15(12), 1509–1519.

Peng, R., Lin, G. and Li, J. (2016) Potential pitfalls of CRISPR/Cas9-mediated genome editing. *FEBS Journal* 283, 1218–1231.

Pillay, M. (1995) Chloroplast DNA similarity of smooth bromegrass with other *Pooid* cereals: Implications for plant breeding. *Crop Science* 35, 869–875.

Pillay, M. and Tripathi, L. (2007) Banana: An overview of breeding and genomics research in *Musa*. In: Kole, C. (ed.) *Genome Mapping and Molecular Breeding in Plants, Volume 4 Fruits and Nuts*. Springer, Heidelberg, Germany, pp. 281–301.

Ray, D.K., Mueller, N.D., West, P.C. and Foley, J.A. (2013) Yield trends are insufficient to double global crop production by 2050. *PLoS ONE* 8, e66428.

Roy, K.R., Smith, J.D., Vonesch, S.C., Lin, G., Tu, C.S., *et al*. (2018). Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nature Biotechnology* 36, 512–520.

Sánchez-León, S., Gil-Humanes, J., Ozuna, C.V., Giménez, M.J., Sousa, C., *et al*. (2018). Low-gluten, nontransgenic wheat engineered with CRISPR/Cas9. *Plant Biotechnology Journal* 16 (4), 902–910.

Savary, S., Ficke, A., Aubertot, J.N. and Hollier, C. (2012) Crop losses due to diseases and their implications for global food production losses and food security. *Food Security* 4, 519–537.

Scheben, A., Wolter, F., Batley, J., Puchta, H. and Edwards, D. (2017) Towards CRISPR/Cas crops – bringing together genomics and genome editing. *New Phytologist* 216, 682–698.

Schreiber, M., Stein, N. and Mascher, M. (2018) Genomic approaches for studying crop evolution. *Genome Biology* 19, 140.

Shi, J., Gao, H., Wang, H., Lafitte, H.R., Archibald, R.L., *et al*. (2017) ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnology Journal* 15, 207–216.

Shi, J., Habben, J.E., Archibald, R.L., Drummond, B.J., Chamberlin, M.A., *et al*. (2015) Overexpression of ARGOS genes modifies plant sensitivity to ethylene, leading to improved drought tolerance in both Arabidopsis and maize. *Plant Physiology* 169, 266–282.

Smith, J., Grizot, S., Arnould, S., Duclert, A., Epinat, J.C., *et al*. (2006) A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Research* 34, e149.

Song, G., Jia, M., Chen, K., Kong, X., Khattak, B., *et al*. (2016) CRISPR/Cas9: A powerful tool for crop genome editing. *The Crop Journal* 4, 75–82.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. and Doudna, J.A. (2014) DNA interrogation by the CRISPR RNA-guide endonuclease Cas9. *Nature* 6, 62–67.

Sun, Y., Jiao, G., Liu, Z., Zhang X, Li, J., *et al*. (2017) Generation of high-amylose rice through CRISPR/Cas9-mediated targeted mutagenesis of starch branching enzymes. *Frontiers in Plant Science* 8, 298.

Unigwe, A.E., Enrico Doria, E., Adebola, P., Gerrano, A, S. and Pillay, M. (2017) Antinutrient analysis of 30 Bambara groundnut (*Vigna subterranea*) accessions in South Africa. *Journal of Crop Improvement* 32, 208–224.

Vasconcelos, M., Datta, K., Oliva, N., Khalekuzzaman, M., Torrizo, L., *et al*. (2003) Enhanced iron and zinc accumulation in transgenic rice with the ferritin gene. *Plant Science* 164, 371–378.

Wang, F., Wang, C., Liu, P., Lei, C., Hao, W., *et al*. (2016) Enhanced rice blast resistance by CRISPR/Cas9-targeted mutagenesis of the ERF transcription factor gene OsERF922. *PLoS ONE* 11(4), e0154027.

Wang, P., Zhang, J., Sun, L., Ma, Y., Xu, J., *et al*. (2018) High efficient multisites genome editing in allotetraploid cotton (*Gossypium hirsutum*) using CRISPR/Cas9 system. *Plant Biotechnology Journal* 16, 137–150.

Wang, S., Yin, L., Tanaka, H., Tanaka, K. and Tsujimoto, H. (2011) Wheat-Aegilops chromosome addition lines showing high iron and zinc contents in grains. *Breeding Science* 61, 189–195.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., *et al*. (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology* 32, 947–951.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., *et al*. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 22, 759–71.

Zhang, G., Lu, T., Miao, W., Sun, L., Tian, M., *et al*. (2017) Genome-wide identification of ABA receptor PYL family and expression analysis of PYLs in response to ABA and osmotic stress in *Gossypium*. *PeerJ* 5, e4126.

Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., *et al*. (2014) The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnology Journal* 12, 797– 807.

Zhong, G.Y. (2001) Genetic issues and pitfalls in transgenic plant breeding. *Euphytica* 118(2), 137–144.

Zsögön, A., Čermák, T., Naves, E.R., Notini, M.M., Edel, K.H., *et al*. (2018) *De novo* domestication of wild tomato using genome editing. *Nature Biotechnology* 36, 1211–1216.

Zuriaga, E., Blanca, J.M., Cordero, L., Sifres, A., William, G., *et al*. (2009) Genetic and bioclimatic variation in *Solanum pimpinellifolium. Genetic Resources and Crop Evolution* 56, 39–51.

5 Epigenome Editing in Crop Improvement

Gurbachan S. Miglani* and Rajveer Singh
Punjab Agricultural University, Ludhiana, India

Introduction

Epigenetics was originally described by Conrad H. Waddington in 1942, while studying differentiation of pluripotent stem cells into different cell types and tissues (Waddington, 1942). To answer the question of what leads the genetically identical stem cells to form phenotypically different cells, performing different functions, without change in DNA sequence, he proposed that 'something is ruling above the DNA' and coined the term 'epigenetics', i.e. 'above genetics'. He explained differential expression of the genome in different cell types. Simply put, there is a layer that sits above the DNA that can influence whether a gene is turned on or off. In general, both genetic and epigenetic variations lead to phenotypic variation. Genetics deals with any change in DNA sequence attributable to mutation and its resulting phenotype, whereas epigenetics deals with DNA methylation and histone modification, leading to modification of gene expression, which may result in a modified phenotype (Fig. 5.1). Epigenetics determines whether a gene will express or not. In all living organisms, epigenetic controls are essential for normal development. Kapazoglou *et al.* (2018) have defined epigenetics as heritable alterations in chromatin architecture that do not involve changes in the underlying DNA sequence but greatly affect gene expression and impact cellular function.

The modifications caused by epigenetics are referred to as epigenetic marks or simply 'epimarks'. A characteristic epimark state at a locus is known as 'epiallele'. In general, there are two important types of epimarks – DNA (cytosine) methylation and histone modifications, which regulate gene expression. These epimarks have a characteristic pattern in different cell types/tissues, known as an 'epigenome'. In comparison with a genome, an epigenome is highly dynamic and sensitive to the environment, as it shows phenotypic plasticity in response to environmental changes, which leads to adaptive phenotype and transgenerational inheritance of such acquired characters (Feng *et al.*, 2012). However, such changes are quite slow and selective. Epigenetic diversity is as important as genetic diversity in crop breeding programmes and it can be harnessed from natural epigenetic variation, also called 'epivariation', which is either already present in germplasm or it can be induced through stress (Gallusci *et al.*, 2017).

Many of the traits, such as flowering, and traits related to biotic and abiotic stresses, are epigenetically regulated (sometimes in complex networks) in relation to the environment. The

* Email: miglanigs1945@gmail.com

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

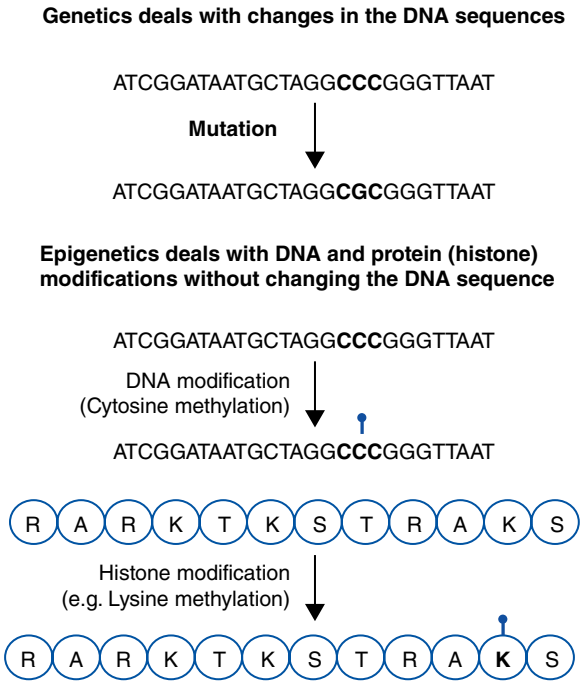


Fig. 5.1. Difference between genetics and epigenetics.

barrier of limited epigenetic variation in plants needs to be overcome through some unconventional technology, such as epigenome editing, to improve crop plants. Thus, after conventional plant breeding, and mutational and targeted genome editing approaches, epigenome editing holds great promise for crop improvement.

Epigenetic Changes

DNA methylation

DNA methylation is an important epigenetic mechanism, wherein DNA methyltransferase (DMT) adds a methyl group to the fifth position of cytosine bases of DNA to form 5-methylcytosine (5^mC). Holliday and Pugh (1975) called the effect of DNA methylation 'repression of gene expression'. DNA methylation pattern in somatic cells is heritable during the process of DNA replication in higher eukaryotes (Jurkowska *et al.*, 2011). It is one of the key processes that induces epigenetic changes in both plants and animals. Most of the genomic regions, including repetitive sequences, heterochromatin and some exonic parts, show a high degree of methylation; in contrast, DNA methylation mostly lacks in promoter regions (Weber and Schubeler, 2007). DNA methylation in mammals is restricted to symmetrical CG sequences (Bird, 2002), whereas plant DNA methylation is found at CpG, CpHpG (both symmetrical sites) and CpHpH (asymmetrical) sites (H could be A, C or T).

In plants, three DMTs have been characterized: methyltransferase 1 (MET1), chromomethylase 3 (CMT3) and domains rearranged methyltransferase 2 (DRM2) (Cao and Jacobsen, 2002). MET1 and CMT3 are maintenance type DMTs; the former is responsible for maintenance of symmetric CG methylation (Jones *et al.*, 2001), whereas the latter relates to the maintenance of DNA methylation at CHG (where H = A, T or C) sites (Lindroth *et al.*, 2001). DRM2 is responsible for *de novo* DNA methylation at all sequences (Pontes *et al.*, 2006); its role in CHH methylation is most prominent, as CHH methylation cannot be maintained, and consequently, it must rely entirely on *de novo* methylation. RNA-directed DNA methylation (RdDM) is an epigenetic process in plants that involves both short and long non-coding RNAs (Matzke *et al.*, 2015). The generation of these RNAs and the

induction of RdDM rely on complex transcriptional machineries comprising two plant-specific, RNA polymerase II (Pol II)-related RNA polymerases known as Pol IV and Pol V, and a number of auxiliary factors (Matzke *et al.*, 2015).

De novo DNA methylation

In plants and animals, there is an epigenetic reprogramming during developmental processes, such as gametogenesis, fertilization and somatic cell reprogramming between sexual generations, to reset epigenetic marks to reduce the risk of perpetuating dangerous epigenetic alleles (Feng *et al.*, 2010). Small/short non-coding RNAs (abbreviated as smRNAs, sncRNAs or sRNAs) and inheritance of DNA and histone marks may also contribute to epigenetic inheritance and reprogramming. The methylated DNA loci are frequently accompanied by small interfering RNAs (siRNAs), which supports the important role of siRNAs in DNA methylation (Lister *et al.*, 2008). Both siRNAs and long non-coding RNAs (lncRNAs) are involved in *de novo* DNA methylation (Wierzbicki *et al.*, 2008).

A model of RdDM in plants was proposed by Bond and Baulcombe (2014). First, Pol IV transcribes the target genomic region into single-stranded RNA, which further gets converted into double-stranded RNA (dsRNA) by RNA-dependent RNA polymerase 2 (RDR2). The dsRNA gets further cleaved into 24-nt smRNA duplexes by dicer-like 3 ribonuclease III enzyme (DCL3) (Xie *et al.*, 2004). The single-stranded 24-nt RNAs are loaded into an AGONAUTE effector protein (AGO4, AGO6 or AGO9) (Havecker *et al.*, 2010) and provide additional target specificity for RdDM. AGO4, for example, interacts with members of the DNA damage response (DDR) complex comprising RDM1, DMS3 and DRD1, which assists in the transcription of nascent scaffold RNAs complementary to the loaded smRNAs produced by another plant-specific, DNA-dependent RNA polymerase, Pol V (Law *et al.*, 2010). The Argonaute (AGO)-loaded smRNAs base pair with the Pol V scaffold RNAs (Wierzbicki *et al.*, 2008, 2009), which then stimulates the recruitment of the *de novo* DNA methyltransferase DRM2 to catalyse *de novo* DNA methylation at the target locus. RdDM, once established, is a robust, self-reinforcing process.

Maintenance of DNA methylation

Maintenance of DNA methylation is a characteristic of individual somatic cell type, which stably transfers the methylation pattern as such into newly formed cells during mitotic divisions. Maintenance of DNA methylation, in a symmetrical context, is governed by DMT that mimeographs the methylation onto the opposite cytosine residue of the newly formed strand (Law *et al.*, 2010). The main enzyme is MET1, which works along with a highly conserved co-factor called p97/valosin-containing protein (VCP)-interacting motif (VIM) in plants and ubiquitin-like, containing plant homeodomain (PHD) and ring finger domains 1 (UHRF1) in mammals, which assists in the maintenance of methylation (Fig. 5.2).

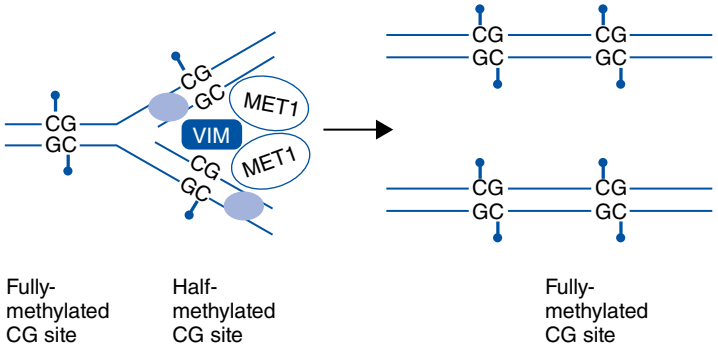


Fig. 5.2. Maintenance of DNA methylation.

Histone modifications

Cellular DNA is firmly packaged in the form of chromatin around a nucleosome octamer consisting of core histone proteins (two copies of each H3, H4, H2A, H2B), around which 147 base pairs of DNA are wrapped. The core histones are largely globular, except for their unstructured N-terminal 'tails'. Histone modifications not only regulate gene expression by holding chromatin either tightly or loosely but also help chromatin remodelling enzymes to reposition nucleosomes through ATP hydrolysis (Bannister and Kouzarides, 2011). Histones are post-translationally modified and there are a large number of distinct types of post-translational modifications (PTMs) found on histones; however, important ones include methylation, acetylation, phosphorylation, ubiquitination, SUMOrylation and poly(ADP)-ribosylation (Kouzarides, 2007). There are histone modifying enzymes which add, remove and read the modification; they are known as histone writers, erasers and readers, respectively (Biswas and Rao, 2018).

Epigenome Modifications in Plants

There is a huge role for epigenetic modifications, including DNA methylation and histone modifications in plants, in the regulation of expression of stress-related genes (Chinnusamy and Zhu, 2009; Levine, 2019). These mechanisms regulate all major aspects of genetic functions, including replication, DNA repair, gene transposition, transcription and cell differentiation. These modifications are tissue-, species-, organelle- and age-specific (Vanyushin and Ashapkin, 2011). Various abiotic and biotic factors influence hormonal fluxes, which, in turn, control DNA methylation (Zhang *et al.*, 2012), resulting in plant adaptation to environmental stresses (Mirouze and Paszkowski, 2011).

Various physiological, developmental and stress stimuli regulate or change DNA methylation status in plants. During stresses, modifications in chromatin and generation of smRNAs have been shown to be involved in transcriptional and post-transcriptional control of gene expression (Angers *et al.*, 2010). Various transposable elements, especially retrotransposons, get activated during environmental stress adaptation. To cope up with various environmental stresses, plants have come up with complex gene regulatory mechanisms, which include DNA methylation, chromatin remodelling and smRNA-based mechanisms to regulate the gene expression in response to the stresses, ultimately leading to climatic adaptations (Sahu *et al.*, 2013).

Transgenerational inheritance of DNA methylation pattern in plants under stress conditions has been reported (Hauser *et al.*, 2011; Feng *et al.*, 2012). Such inherited epigenetic plasticity plays an important role in plants' immediate response to and establishment of long-term adaption under stress. However, the mechanism of transgenerational inheritance remains unclear.

Epigenome Diversity Useful for Crop Breeding

In relation to crop improvement, plant breeders have considered genetic diversity or genetic variation as the most important contributor of phenotypic variation and thus have neglected the yet unexplored wealth of epigenetic variation present in the population for a number of traits. Like genetic variation, epigenetic variation is equally important, as it can also be harnessed for crop improvement. Real evidence showing the importance of epigenetic variation came when genome-wide association studies failed to explain considerably important heritable variation present in the natural population of species not having much genetic variation (Johannes *et al.*, 2009). Differential DNA methylation in the natural population of a plant species may act as substantial natural epivariation for important agronomic traits, showing an important role in traditional approaches to plant breeding and improvement (He *et al.*, 2011). More than 99% of the methylome is conserved within a species but still there is a range (hundreds to thousands) of differentially methylated regions (DMRs) present between accessions (Li *et al.*, 2015). Whole-genome DNA methylation-profiling studies have revealed substantial levels of natural variation among accessions of maize (*Zea mays* L.), thale cress (*Arabidopsis thaliana* (L.) Heynh.), soybean (*Glycine max* (L.) Merr.), rice (*Oryza sativa* L.) and *Brachypodium distachyon* (L.) P. Beauv. (Nathan and Robert, 2017). The majority

of DMRs do not exhibit changes in expression of nearby genes (Eichten *et al.*, 2013; Schmitz *et al.*, 2013a). However, for ~10–20% of DMRs, there is a negative association between methylation and gene expression, suggesting that a subset of methylation variation has the potential to influence phenotype. Altered levels of DNA methylation are more effective for genes that exhibit qualitative differences in on–off expression than for genes showing quantitative differences in expression (Li *et al.*, 2015). The natural variation of DNA methylation and its potential to influence gene expression and plant traits can be easily highlighted through genome-wide analyses of methylomes and transcriptomes.

A pool of epigenetic variation or epigenetic diversity linked to specific agronomic traits needs to be exploited for crop improvement. Although epigenetic diversity is promising for designing phenotypically diverse crop plants, the number of known natural epimutations in crop plants is limited (Iwasaki and Paszkowski, 2014). However, in addition to natural epigenetic diversity, several new ways are now available to induce epigenetic variation (Gallusci *et al.*, 2017).

Non-targeted epigenetic diversity

Non-targeted epigenetic diversity refers to the random epigenetic variation present in the genome. Varied phenotypes were seen in *A. thaliana* by generating crosses between *ddm1* hypomethylated mutants with isogenic wild-type lines and the resultant progeny, called epigenetic recombinant inbred lines (EpiRILs) (Johannes *et al.*, 2009). The behaviour of EpiRILs is equivalent to recombinant inbred lines (RILs) produced in breeding programmes, but their effect is attributable to DNA methylation differences and resultant gene expression in individuals. Generating such mutants in other crops may be time consuming, but utilizing pharmacological inhibitors (e.g. azacytidine or zebularine) of DNA methyltransferase is an alternative approach to hypomethylate the parental epigenome (Akimoto *et al.*, 2007). Selection of isogenic lines bearing epigenetic components controlling agronomic traits could be incorporated in breeding schemes (Hauben *et al.*, 2009). Natural epigenetic variation can be exploited and heritable natural epialleles are an interesting option for crop improvement, although only a few have been described. Extensive epigenomic studies are required to identify new epialleles for their possible association with traits of interest. Hypomethylated genomes and epiRILs can also be used to identify DMRs related to the trait(s) of interest (Schmitz *et al.*, 2013b). In short, epigenetic modifications induced by stresses provide an alternative source of natural epigenetic diversity. Stress can be monitored to experimentally generate stable epigenetic variants, such as those obtained using *in vitro* culture (Chen *et al.*, 2015).

Targeted epigenetic diversity

To induce target-specific epigenetic variations, *a priori* knowledge about the epigenetic state of target sequence is required. This can be achieved through target-specific epigenome editing methods that have been successfully used in animal systems to either add or remove the selected chromatin modifications at defined loci through various approaches (Thakore *et al.*, 2016). Epigenome editing tools are now being developed and perfected in plants. Epigenome editing requires efficient and target-specific DNA targeting machinery, transformation methods and post-experimental analysis of target site. The inheritance of induced targeted epigenetic variation has been reported up to third generation in *A. thaliana* (Gallego-Bartolomé *et al.*, 2018) and up to six or seven generations in rice (Akimoto *et al.*, 2007), but its inheritance is stable in vegetatively propagated crops (Kasai *et al.*, 2016). The heritability of induced epigenetic modifications in sexually reproduced crops is still a puzzle. This approach for crop breeding may be limited by a lack of knowledge about inheritance of the induced epigenetic modifications (Kungulovski *et al.*, 2015). Thus, the stability and heritability of the induced epigenetic modifications would have to be critically evaluated (Kasai and Kanazawa, 2013).

Genome Editing to Epigenome Editing

'Genome editing' refers to any procedure with which a target-specific DNA sequence can be altered. The altered DNA sequence may comprise an insertion/deletion or substitution of at least a single nucleotide. The edited DNA sequence may

encode a modified gene product (e.g. a protein with an altered amino acid (aa) sequence, a non-coding RNA with an altered nucleotide sequence), may provide a modified function (e.g. altered promoter function, altered enhancer function) or may fail to give rise to a gene product (Miglani, 2017, 2019). Genome editing is based on producing a target-specific double-stranded break (DSB) in DNA. The genome editing toolbox comprises many engineered molecular scissors for inducing DSBs at specific genomic loci. These molecular scissors include customized, site-specific nucleases (SSNs), such as meganucleases (MNs), also known as homing nucleases (HNs), zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and RNA-guided nuclease (RGN) systems. The most popular example of RGN is the clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (CRISPR/Cas9) technology. Two major DNA repair pathways have been exploited for genome editing: non-homologous end-joining (NHEJ) and homology-directed repair (HDR), which are endogenous to both prokaryotes and eukaryotes (Puchta, 2005).

Research on genome editing in crops started only 6–7 years back and is still in its infancy. Now there is an emergence of global perspective that modulating gene expression will be more important for generating future crops capable of coping with environmental stresses. Most of the organisms, especially plants, use such regulatory mechanisms the most, as they are immobile and have to cope with changing external conditions. The two important pillars of fundamental epigenetics, viz., DNA methylation and histone modification, which play a crucial role in controlling genomic functions in relation to stresses, have now been well studied in plant species, such as rice, *Arabidopsis*, barley, tobacco, tomato, soybean and maize (Saraswat *et al*., 2017). With the availability of epigenome maps, including both DNA methylation and histone tags, in various crop species, it may now be possible to predict the regulatory role of epigenetic pattern across particular genomic regions, especially promoters and enhancers. The information thus obtained can be even more useful when such comparative studies are extended to large natural populations for the analysis of the variation in epigenetic patterns across such regions. In certain crop species that lack a significant level of epigenetic variation, variation can be generated through the currently available epigenome-editing technology, also called targeted chromatin rewriting. Targeted epigenome editing was first indicated by Choo *et al*. (1994) when they targeted a nine-base pair region of a *breakpoint cluster region-Abelson* (*BCR-ABL*) fusion oncogene using a zinc-finger protein. The successful application of this technology was realized in human embryonic kidney cells (HEK293) when it was reported that genes could be repressed by a histone methyltransferase (HMT) functional catalytic core fused with a DNA-binding domain of a zinc-finger protein (ZFP) (Snowden *et al*., 2002).

The term epigenome editing refers to the procedures of chromatin engineering, in which the epigenome is modified at specific sites. Unlike genome editing, epigenome editing does not involve any change in nucleotide sequences; rather it involves presenting the DNA sequence to DNA-binding protein domains that influence DNA function in the presence of an associated effector domain (e.g. DNA methyltransferase, histone acetylase). With the advent of precise epigenome editing tools, it is now experimentally possible to modify individual chromatin marks at specific user-defined sites (Köferle *et al*., 2015). One can determine the biological role of epigenetic marks across gene regulatory sequences by targeting them through site-specific epigenome editing tools and observe and validate the phenotypic effects of such modifications. Before resorting to the use of epigenome editing technology for crop improvement, the biological effect of each and every epimark change on plant phenotype must be clearly elucidated.

Why Go for Epigenome Editing?

Various valuable plant traits, such as seed dormancy, vernalization, flowering time, disease resistance, micronutrient levels, pigmentation, fruit ripening, yield attributes, energy homeostasis and secondary metabolism, controlled through transcriptional networks, are under the strict regulation of various epigenetic mechanisms, such as DNA methylation, non-coding RNAs, chromatin remodelling and histone modification, which add a level of transcriptional regulation

(https://www.mdpi.com/journal/agronomy/special_issues/epigenetic_mechanisms_in_crop_plants). Besides these, complex traits, such as abiotic (Sahu *et al.*, 2013) and biotic stress tolerance (Espinas *et al.*, 2016), are also under epigenetic control.

Although sufficient epigenomic variation may be available in various pre-cultivated and wild accessions, its extent is limited because of tight regulation of maximum common regions (Li *et al.*, 2015). In all the presently available accessions, in which the expression of certain genes (e.g. those for disease resistance) is dormant, using epigenome editing, expression of such genes can be reactivated (Akimoto *et al.*, 2007). This approach may thus serve as a new source of stable phenotypic variation, which otherwise could not be harnessed through conventional breeding and mutagenesis. Although naturally induced epigenetic variation attributable to environmental factors could be utilized, it is quite rare, slow and applicable only to certain traits. Thus, there is a need for developing an approach to break such boundaries of epigenetic variation and utilize the same for organismal improvement. Moreover, certain endogenous regulatory elements existing in present crop species are not able to modulate their expression according to environmental changes, so epigenome editing may help in this regard. Most of the important plant traits are regulated epigenetically but some species lack epigenomic variation in some populations for a given trait. Moreover, the environment has a more direct effect on the epigenome compared with the genome. Epigenome changes are also heritable; and, most importantly, epigenome editing causes no change in the native genome sequence.

Strategies of Epigenome Editing

Kanherkar *et al.* (2014) have explained modelistic approaches to alter the epigenome. Accordingly, an epigenetic factor (any molecular inhibitor) can change a particular epigenetic state through a direct or an indirect mechanism (Fig. 5.3). A direct effect could be produced in two ways – Type 1

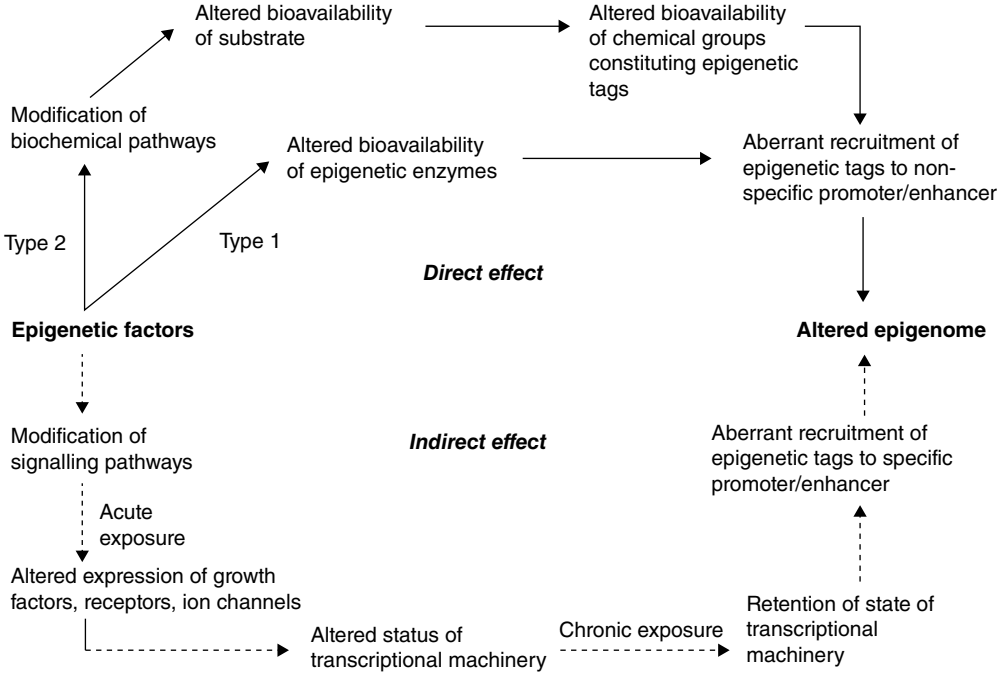


Fig. 5.3. Strategies to alter the epigenome. (From Kanherkar *et al*., 2014 under the terms of the creative commons attribution license, CC BY.)

and Type 2. In Type 1, an epigenetic factor directly binds to epigenetic enzymes and alters their normal function, either damaging them in some way or by upregulating them, which results in aberrant recruitment of epigenetic tags to promoters and enhancers on a genome-wide scale. Such a direct effect results in random alteration in the epigenome. An example of Type 1 direct effects is the antihypersensitive hydralazine, which inhibits DNA methylation (Chavez-Blanco *et al.*, 2006). Type 2 direct effects occur when an epigenetic factor causes a change in a biochemical process by targeting the availability of a substrate, intermediate, by-product or any other metabolite participating in the biochemical pathway, which is used to make up epigenetic tags (e.g. acetate) (Kanherkar *et al.*, 2014). This in turn leads to a non-specific modification of the epigenome. The second way to cause epigenetic changes is through indirect mechanisms. In this case, a small-time exposure of a target to a factor influences cellular signalling pathways, which leads to altered expression of growth factors, receptors and ion channels, and in turn alters transcription factor activity at gene promoters (Kanherkar *et al.*, 2014). However, with chronic exposure, the regulatory machinery may get affected leading to altered gene expression by actually recruiting or repelling epigenetic enzymes to/from the associated chromatin, resulting in the addition or removal of epigenetic tags. Due to recent developments in epigenome editing tools, the targeted recruitment of epigenetic tags is now possible over regulatory sequences (e.g. promoters, enhancers, insulators). Such epigenome editing tools include DNA binding proteins fused with effector molecule.

Prerequisites for Epigenome Editing

If the objective is to induce global changes in the epigenomic landscape, then random methods of epigenome editing are used, but if the objective is to induce modification(s) at a specific locus or a region (such as promoter, enhancer, silencer and insulator), precise methods of epigenome editing are used (Thakore *et al.*, 2016). In the latter case, epigenetic profiling of the target region related to target trait is necessary. Epigenetic profiling (Fig. 5.4) includes determining the methylation pattern, histone tags, expression profiling of epigenetic modifiers, such as sncRNAs and lncRNAs, which regulate the expression of the target locus/region or gene. Epigenetic tags across the target region can be directly modified by targeted erasing and rewriting the epimarks or indirectly by targeting the epigenetic modifiers. However, critical profiling and knowledge of such modifiers is quite necessary whether they govern related or unrelated traits.

Epigenetic profiling is highly critical for assigning the role of the epigenetic pattern across the target region to its resulting phenotype. In other words, it is the co-study of epigenetic patterns or epimarks and the related phenotype, depending upon the type of regulation (negative or positive) imparted by such epigenetic marks across a particular target region. Thus, linking the epigenetic profile across a particular target region with its existing phenotype/gene expression

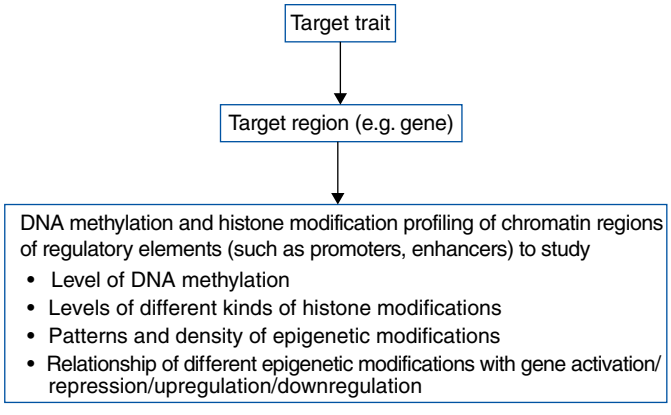


Fig. 5.4. Epigenetic profiling of target region.

(Romanowaska and Joshi, 2019) is the first step in setting up an epigenome editing experiment. Profiling of epigenetic tags across the target region is not sufficient. Success of an epigenome editing experiment could be greater if mapping and expression profiling of endogenous epigenetic modifiers were done, e.g. finding out which ncRNAs are responsible for creating such epigenetic patterns across the target region (He *et al.*, 2011). Targeting such molecules originating from genomic regions will be a permanent solution to the stable inheritance of epigenetic modification; otherwise, inheritance of induced epigenetic modification in sexually reproducing crops will still be doubtful. The last, but not the least, requirement for efficient epigenome editing is the precision of the method with little off-target effects, so as to have the desired modification in the parental background (Yang *et al.*, 2018).

Epigenetic Profiling of Target Region Related to Target Trait

Epigenetic profiling of target trait to target the regulatory region is not simply determining the epigenetic pattern but also determining the effectors or modifiers that cause such epigenetic patterns across such regions (He *et al.*, 2011). Targeting epigenetic modifiers (such as short non-coding RNAs, long non-coding RNAs) originating from genomic regions are the ultimate targets for a permanent epigenome editing programme. Usually these modifiers have the same or nearby position to the region they target for epigenetic modification. Thus, profiling of both epigenetic patterns across the target region and their respective effectors/modifiers is crucial. There are many technologies available for high-throughput profiling of epigenetic marks and modifiers that cause such modifications across the target region. However, in many complex traits, such as fruit ripening and other defence-related responses (biotic and abiotic), where complex genetic networks, including multiple epigenetic factors for either a related or unrelated trait work together (Li *et al.*, 2018a), their high-throughput sequencing and differential expression profiling of the related transcripts, including genes and non-coding transcripts, is necessary. Presently, many transcriptome and epigenome profiling techniques are in use, e.g. genome-tilling microarray, high-throughput sequencing, bisulfite sequencing, histone mapping and smRNA profiling (He *et al.*, 2011). Whole-genome methylation maps, called methylomes, are available for different tissues in different crops (Ji *et al.*, 2015), which provide methylation patterns across the regulatory and genic regions. From such methylation data, regulatory or genic regions can be assessed to see whether these genomic regions are heavily methylated or hypomethylated. Such observations can then be linked with their expression and this information should enable researchers to target such regions for modulation of gene expression.

Recently, whole genome bisulfite sequencing (WGBS) data, combined with METHimpute, have enabled the development of whole methylomes by presenting the methylated cytosines in the genome, irrespective of coverage (Taudt *et al.*, 2018). Similarly, for histone modification, next-generation sequencing (ChIP-seq) data have been integrated with chromstaR, an algorithm, to enable the computational interpretation of the combined effect of different epigenetic states across an unpredictable number of conditions, such as different tissues or individuals.

Non-coding RNAs and their Genome-wide Profiling in Plants

The ncRNAs are key players in gene regulation with zero potential to form proteins (low protein-coding potential); thus, they are produced as mRNA from DNA but are not translated into proteins. The ncRNAs, based upon their length, have been divided into three classes: small non-coding RNAs (sncRNAs) (18–30 nt), medium-sized ncRNAs (mncRNAs) (31–200 nt) and lncRNAs (>200 nt). The siRNAs have a direct role in regulation of target mRNAs and chromatin; they are best characterized for their role in transcriptional gene silencing (TGS) (Xu *et al.*, 2013) apart from post-transcriptional gene silencing (PTGS). Mostly siRNAs are involved in PTGS, but hc-siRNAs are involved in TGS (Ku *et al.*, 2015). The ncRNAs can follow different mechanisms by which they can interact with genes to upregulate or downregulate gene expression, to inhibit protein synthesis, or guide methylation (Amaral *et al.*, 2008; Collins and Chen, 2009; Collins and Penny, 2009).

Identifying the Role of ncRNAs in Plants

The prerequisite for the use of ncRNAs (siRNAs and lncRNAs) in genome editing is their identification. Trait-specific smRNAs and lncRNAs can be extracted by comparing the RNA profiles between treated and untreated plants through high-throughput differential expression and sequencing (transcriptome study). Differential expression profiling of coding (genes) and ncRNAs through quantitative real-time PCR (qPCR) is the key step for assigning a role to them in a particular stress environment. Therefore, the identification of sRNAs involved in the gene regulation network requires reliable computational tools. Methods for the identification and assigning role to siRNAs and lncRNA are discussed below.

siRNAs identification

In plants, the massively parallel signature sequencing (MPSS) approach was first performed for genome-wide profiling of siRNAs (He *et al.*, 2011). With the advent of next-generation high-throughput sequencing technologies, high-resolution ncRNA maps were obtained using the 454 and Illumina sequencing systems, which are well suited for ncRNA discovery and profiling. For smRNA mapping, all sequencing libraries are processed in the same manner. The 3′ adapter trimming is performed for perfect matching. The sequences not overlapping with the adaptor on the 3′ end are discarded. The remaining reads are then subjected to size selection and can be mapped against the genome. Based on such methods, Hardcastle *et al.* (2018) produced an smRNA locus map in *A. thaliana*. Through such smRNA locus maps, epigenetic regulation of such molecules can be studied.

lncRNAs identification

lncRNA related to target trait can only be characterized through differential screening in different genotypes, e.g. stress-tolerant and stress-susceptible genotypes. Under stress conditions, these genotypes will give different responses in the form of tolerance and susceptibility, which is attributable to the differential expression of certain lncRNAs in both the genotypes (Fig. 5.5). The lncRNAs will be differentially expressed (i.e., some will be upregulated while others will

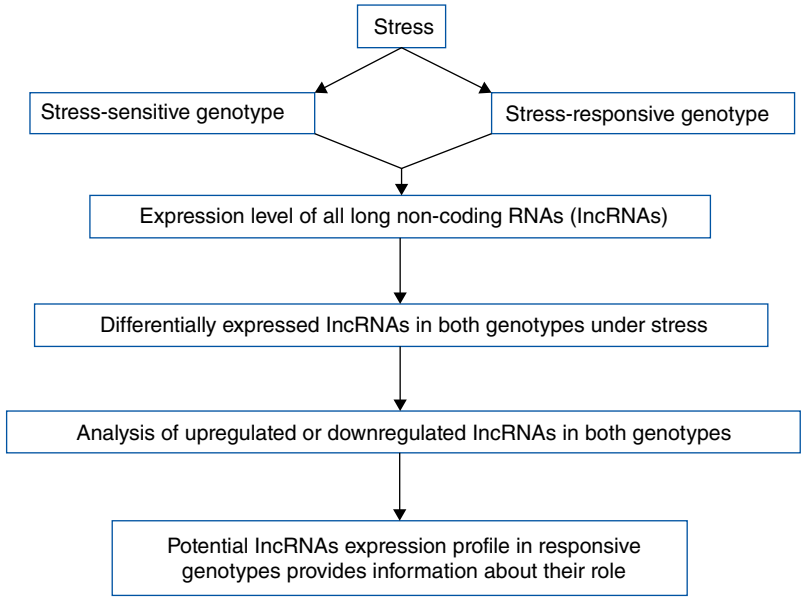


Fig. 5.5. Differential expression profiling of long non-coding RNAs (lncRNAs).

be downregulated in both the genotypes). The resultant upregulated or downregulated lncRNAs in the tolerant genotype informs the role of that particular lncRNA under stress conditions. For instance, Li *et al*. (2018c) have identified tobacco lncRNAs responsive to root-knot nematode stress. The identification of suitable sized transcripts and their non-coding potential are two main steps in identifying lncRNAs (Fig. 5.6). First, the high-throughput sequencing data or tiling microarrays data are used to identify lncRNA transcript units. The regenerated data are assembled and annotated for protein-coding transcripts; as lncRNAs are non-coding transcripts, all the transcripts with protein-coding potential are rejected. Second, the codon potentials of these transcript units are calculated on the basis of codon statistics and similarity to known protein-coding sequences. The transcripts that have zero protein-coding potential are usually selected (Wang *et al*., 2017). Several tools have been designed and used to evaluate coding potential, e.g. coding potential calculation (CPC) (Kong *et al*., 2007). CPC tools have been developed on the basis of support vector machine (SVM, which is a machine-learning algorithm that can be used for both regression and classification challenges). However, it is mostly used for classification. A number of other databases are available for predicting lncRNAs (Wang *et al*., 2017); however, the most recent one is pLncPRO (Singh *et al*., 2017), which has better prediction accuracy as compared to other existing tools and is particularly well suited to plants.

Epigenome Editing Approaches

To date, several methods have been exploited for producing desired epigenetic modifications in plant genomes. Depending upon the objective of research, the epigenome can be modified either randomly or at a target-specific locus. Here we discuss the methods that have so far been developed to induce epigenomic variation, including both random and precise methods of epigenome editing.

Random epigenome editing methods

For random epigenome editing, three methods have thus far been employed: (i) antimetabolite inhibitors; (ii) tissue culture-based approach; and (iii) overexpression of effector molecules. These methods are discussed here briefly.

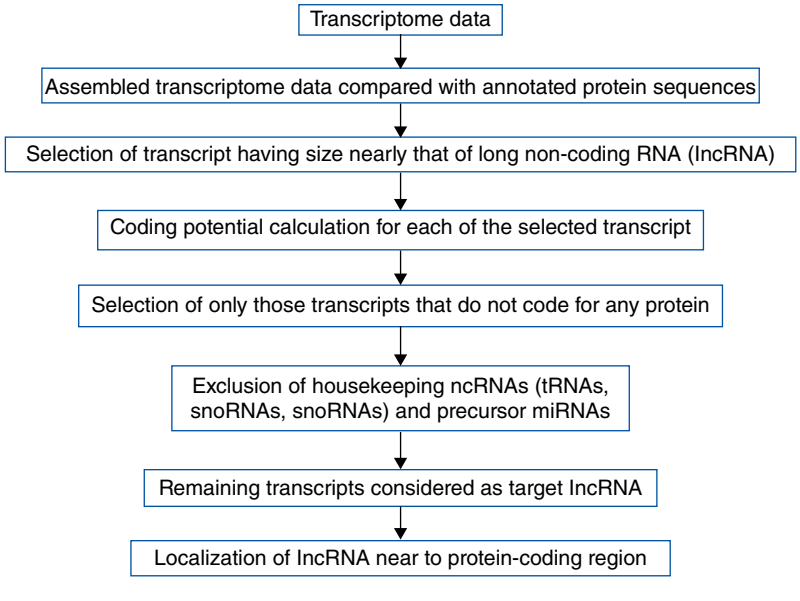


Fig. 5.6. Identification and mapping of long non-coding RNAs (lncRNAs) over target region.

Epigenome editing through antimetabolite inhibitors

Epigenome editing in crops started with a conventional non-precise (totally random) method utilizing nucleoside inhibitors, such as 5-azacytidine (5azaC) and 5-aza-2'-deoxycytidine (5azadC). The nucleoside inhibitors are incorporated into DNA to inhibit or reduce the function of DNA methyltransferase. Nucleoside antimetabolites have been employed in *Nicotiana tabacum* to activate the silent genes (Zhu *et al.*, 1991). This led to widespread use of 5azaC and 5azadC to induce hypomethylation in certain gene regions to activate the associated genes. The use of 5azadC in rice has been demonstrated, where six hypomethylated fragments (HMFs) were differentially expressed in a selected mutant line-2 compared to wild type (Akimoto *et al.*, 2007). Line-2 showed a clear marker phenotype of dwarfism, which was stably inherited by the progeny across nine generations. Out of six HMFs, HMF2 encodes retrotransposon gag-pol polyprotein, which showed a complete erasure of 5-methylcytosine ($5^{m}C$). Similarly, HMF5 encodes putative Xa21-like protein called Xa21G, which resulted in the acquisition of disease resistance in rice. Thus, treatment of germinated seeds with 5azadC had induced phenotypic changes evident at maturity; such changes were stably inherited by the progeny.

Epigenome editing through tissue culture

The *in vitro*-regenerated plants possess built-in variation in them. This variation is attributed to different conditions in artificial microenvironments, which plant cells/tissues face (Us-Camas *et al.*, 2014). Cellular growth and organogenesis depend upon several genetic and epigenetic factors and coordination among them.

Recent studies have enhanced our knowledge about the epigenetic role and its dynamics (including both DNA methylation and histone modification) in cellular differentiation and cell fate changes (Lee and Seo, 2018). During callus formation, not only genetic variation but also significant epigenetic variation has been observed among and within species because of loss of methylation (hypomethylation), e.g. rice callus. Such methylation changes are largely around gene promoter sequences, which result in liberal expression of protein-coding genes and, thus, certain biological processes. The resultant regenerants, known as somaclonal variants, show variable inheritance of acquired epigenetic changes. Although DMTs check activity of the transposable elements (TEs), during callus formation, most of the TEs may also get hypomethylated and get activated, leading to large-scale transposition that results in progressive inactivation of host genes. Several tissue culture-based locus-specific epialleles have been characterized (Ong-Abdullah *et al.*, 2015). Through tissue culture, genome-wide changes in DNA methylation have been achieved in *Arabidopsis*, rice and maize (Stelpflug *et al.*, 2014). Some epialleles can generate new phenotypes; for instance, the *Karma* epiallele in oil palm is associated with loss of CHG (where H = A, T or C) methylation at a transposable element (TE) nested in a gene (Ong-Abdullah *et al.*, 2015), resulting in the production of aberrant transcripts. Apart from methylation, global histone modifications, including active marks, such as H3 and H4 acetylation, H3K4me3, H3K36me3 and H2Aub, tend to be increased in callus cells, whereas H3K9me2/me3 and H3K27me2/me3, as repressive marks, are globally decreased in dedifferentiated cells.

Epigenetic modifications of specific biological processes have been observed during callus formation (Lee and Seo, 2018). Various examples of such biological processes include acquisition of cell fate changes related to root identity, elimination of leaf characteristics, root stem cell patterning, cell cycle re-entry and progression, and reacquisition of embryonic characteristics.

Tissue culture-derived epialleles follow Mendelian inheritance, which is well evidenced in rice and maize (Stroud *et al.*, 2013; Stelpflug *et al.*, 2014). Moreover, stability of DNA methylation patterns through tissue culture and *ex vitro* acclimatization have been documented in myrtle (*Myrtus communis*) (Parra *et al.*, 2001), agave (De-la-Peña *et al.*, 2012) and almond (Martins *et al.*, 2004). This stability may help in propagating elite individuals over transgenic lines.

The other key factors that play a role during plant tissue culture (PTC) are micro-RNAs (miRNAs) (Chen *et al.*, 2013). The nuclear-localized

miRNAs can act as transcriptional gene silencing (TGS) factors and are thus, associated with epigenetic modification.

The *in vitro* conditions constituting plant growth regulators and stress-based hormone signalling (mechanical or biological) induce miRNA production (Liang *et al.*, 2012). Being small molecules, miRNAs potentially show cell-to-cell communication and thus can be mobilized through grafting tissues between cells to carry the signal to display the epigenetic modifications to other tissues (McGarry and Kragler, 2013). The mobilization of miRNAs, from the adult explant to the juvenile explant, promotes rejuvenation of tissues. The main advantage of graft-induced miRNA/siRNA mobilization is that it is free from the regulatory frames applicable to transgenic crops (Kasai *et al.*, 2016). During the acclimation process, the presence of miR172 promotes SQUAMOSA promoter-binding protein-like (SPL) expression and maturation of tissues to cope with greenhouse conditions. miR156 and miR172 play important roles during micropropagation (Us-Camas *et al.*, 2014). During the globular stage of embryogenesis, miR156, 164, 390 and 397 regulate a development-related NAC (NAM, ATAF1-2, CUC2) family transcription factor and miRNAs 166, 167 and 398 regulate the cotyledon stage by promoting ARF and GH3 gene expression and, in turn, regulate the free 3-indolacetic acid in the cells for the promotion of the SCF-TIR1 complex function (Siddiqui *et al.*, 2018). However, there is no information available regarding miRNAs involved in the 'heart stage'.

In the future, miRNA mobility in tissue culture may provide a strategy for propagating recalcitrant species, such as trees to promote rejuvenation. There is speculation that plant cells can also secrete miRNAs into the media, as animal cells do (Wang *et al.*, 2010). If so, it would be fascinating to see the effect of smRNAs in culture media for recalcitrant plants.

Overexpression of epigenetic effector/modifiers

Another potent method of enhancing global epigenetic variation in crops for rapid selection of epiRILs involves the transgenic overexpression of epigenetic effector/modifiers (either methylase or demethylase) in plant genomes. Methylation stably repressed a few genes throughout the soma and represents an untapped source of hidden genetic variation if transcriptionally reactivated, as revealed from pioneering studies in the model plant *A. thaliana* (Cortijo *et al.*, 2014). However, this new source of genetic variation was realized by creating epiRILs from crosses between a normal methylated wild-type individual and a hypomethylated mutant (*met*1). The morphological variation of epiRILs revealed extensive hidden genetic variation in plant genomes that can be observed because of the expression of newly unmethylated regions. Unfortunately, the *A. thaliana met1* mutant often shows lethality in crops (Hu *et al.*, 2014; Li *et al.*, 2014b). So, Ji *et al.* (2018) proposed a new method of inducing global hypomethylation in *A. thaliana* by transgenic overexpression of a human ten-eleven translocation methylcytosine dioxygenase 1 catalytic domain (hTET1cd), a demethylase enzyme under the control of the CaMV35S promoter. WGBS analysis of two independently derived events (35S:TET1-1 and 35S:TET1-2) revealed that more than one-third of DMRs were located in intergenic sequences, more than half overlapped with genes, and a small fraction was located in promoter regions (≤1 kb upstream of a gene). The extent of CG methylation loss caused by hTETcd expression is lower in *35S:TET1-1* (9.9 Mb) and *35S:TET1-2* (18.0 Mb) than in *met1* (31.8 Mb). In the transgenic plants that were used for WGBS, Ji *et al.* (2018) observed a delay in the developmental transition from vegetative to reproductive growth. They also noted that late flowering phenotype was associated with the demethylation of the *FLOWERING WAGENINGEN* (*FWA*) locus. A closer inspection of the DNA methylation status of this locus revealed that DNA methylation was completely abolished, as was methylation at adjacent CHG and CHH sites.

Precise Epigenome Editing Methods

The site-specific methods of epigenome editing are classified into these three categories: (i) targeting of siRNAs; (ii) targeting of lncRNAs; and (iii) engineered DNA-binding proteins fused with effector molecules. The development of these precise epigenome editing methods has raised

researchers' hopes of using epigenetic approaches for crop breeding programmes.

Targeting of siRNAs

siRNAs are well-known gene-silencing molecules that follow the RdDM pathway (Aufsatz *et al.*, 2002; Henderson and Jacobsen, 2007). Posing as an siRNA molecule showing homologous pairing to the structural element, such as promoter sequence of a gene, can induce methylation across such a region, leading to transcriptional gene silencing (TGS) of that gene (Matzke *et al.*, 2009). Therefore, improvement of clonally propagated crop species (such as potato, sugarcane, fruit trees) through siRNAs would be a particularly unique and promising approach (Kasai *et al.*, 2016). Many studies have demonstrated that siRNAs derived from transgenes can move across the graft union between a scion and a rootstock [e.g. *Nicotiana benthamiana* (Bai *et al.*, 2010); *A. thaliana* (Molnar *et al.*, 2010)].

Kasai *et al.* (2016) demonstrated TGS through the graft union method by mobilizing siRNA derived from the transgenic *N. benthamiana* under companion cell-specific promoter (e.g. Commelina yellow mottle virus promoter) Co35SpIR to target the 35S:*GFP* locus (promoter) in transgenic potato. For a comparative study, a PTGS starter (CoGFPIR) was used to target an exon of 35S:GFP transgene to distinguish it from TGS. The hetero-grafted plants were grown aseptically in a tissue-culture vessel, and after lateral growth, plants were placed on micro-tuber (MT)-induction medium. The change in GFP expression was studied in MTs of two grafted plants, A and B. One tuber from the grafted plant A exhibited almost the same GFP expression as the 35S:*GFP* control; two MTs from the grafted plant B showed a decreased expression. Furthermore, bisulfite sequencing showed clearly high methylation of the target region. In the case of PTGS, GFP transcripts were reduced, but the methylation level remained the same as that of the control. The granule-bound starch synthase 1 (*GBSS1*) gene promoter in potato was targeted by mobilizing artificial siRNA molecules from transgenic tobacco into potato root stock through grafting (Kasai *et al.*, 2016). The resultant plants had waxy-type potato starch and a smooth pulpy texture with low amylose and high amylopectin content. These plants had a high-quality taste, with high viscosity, and were less retrograde in comparison with regular potato starch.

Targeting of lncRNAs

lncRNAs are the important epigenetic players, which regulate gene expression by interacting with DNA and induce differential methylation resulting in up- and down-regulation of gene expression (Heo *et al.*, 2013). Large numbers of lncRNAs have been identified in plants, such as *Arabidopsis* (Zhu and Wang, 2012; Wang *et al.*, 2014), maize (Li *et al.*, 2014a; Huanca-Mamani *et al.*, 2018), rice (Li *et al.*, 2007), wheat (Xin *et al.*, 2011) and tomato (Zhu *et al.*, 2015). Depending upon their role, whether they act as positive or negative regulators of traits, they can be either knocked out or overexpressed.

Zhu *et al.* (2015) reported ripening-related lncRNA1459 in tomato. lncRNA1459 is a potential element in fruit ripening and showed a relatively high expression level in fruit. Li *et al.* (2018b) validated its function by designing a CRISPR/Cas9-based construct to target lncRNA1459, which was transformed into *Solanum lycopersicum* cv. Ailsa Craig. Three heterozygous mutants – *CR-lncRNA1459-32*, *CR-lncRNA1459-8* and *CR-lncRNA1459-21*, with small indels – were compared with the wild type; mutant *CR-lncRNA1459-32* did not show any delay in ripening, whereas the other two mutants (*CR-lncRNA1459-8* and *CRlncRNA1459-21*) showed delayed ripening until 35 days post-anthesis (DPA). Samples of fruits of the wild type and transgenic mutants at the 35 DPA stage were used to extract total RNA for qPCR analysis. The qPCR analysis revealed that transcript levels of lncRNA1459 were distinctly reduced in transgenic *CR-lncRNA1459* mutants compared with the wild-type fruits. RNA-seq analysis discovered thousands of genes, including coding genes and lncRNAs, which were specifically differentially expressed in *CR-lncRNA1459-8-3* and *CR-lncRNA1459-21-1* relative to the wild type. Knocking out lncRNA1459 results in downregulation of demethylase SDML2, which regulates tomato-ripening-related genes, thus resulting in the delayed-ripening phenotype (Li *et al.*, 2018b).

Wang *et al.* (2018) identified a lncRNA transcript, named LAIR, which is an antisense intergenic RNA in leucine-rich repeat receptor kinase (LRK) gene cluster. LRK genes are associated with grain yield in rice. Chromatin modifications related to LRK genes indicated H3K4me3 and H4K16ac to be enriched in transcriptionally active genes and LAIR was responsible for the modifications.

Engineered DNA-binding domains fused with effector molecules

The fundamental principle of precise epigenome-editing technology is based on the fusion of a DNA-recognition/binding domain (DBD) with a catalytic effector domain (ED) of a chromatin-modifying enzyme to generate targeted epieffectors (Kungulovski and Jeltsch, 2016). A short peptide serves as a link between DBD and a catalytic ED (Fig. 5.7) in all the genome- as well as epigenome editing tools. Here we will confine our discussion only to epigenome editing tools.

DNA-binding proteins without nuclease activity, such as ZFPs, transcription activator-like effectors (TALEs) or ribonucleoproteins [e.g. dead Cas9 (dCas9) complexed with single-guide RNA (sgRNA), act as DBDs (Brocken *et al.*, 2018)]. A DBD binds to a specific DNA sequence and helps to recruit a functional catalytic domain to the defined target loci in the genome, where it can sit and modify the chromatin state, and thereby modulate the epigenetic state of DNA or histones, depending upon the specificity of the DBD and nature of the ED. In a genome/epigenome editing tool, DBD is either directly fused to ED or joined indirectly to ED through a linker peptide.

Targeted specific upregulation and downregulation of gene expression can be achieved by posing artificial transcription factors (ATFs) to regulatory elements (promoters, enhancers) bearing chromatin regions, and they are actually being used with presently existing three epigenome-editing platforms. ATFs consist of an ED coupled to a sequence-specific DBD (Heiderscheit *et al.*, 2018). ED affects transcription by recruiting or blocking transcriptional machinery. ATFs have been designed for reprogramming cell fate by modulating certain gene expressions (Heiderscheit *et al.*, 2018), which otherwise cannot be possible through naturally existing transcription factors.

Zinc finger protein-based epigenome editing tools

ZFPs consist of modular zinc finger C_2H_2 domains (20–30 amino acids containing α-helix and two β-sheets coordinated by zinc ion), which are common types of DNA-binding motifs found in eukaryotes and are naturally occurring transcription factors. During single zinc finger-binding, the amino acids of alpha-helix at position 1, 3 and 6 recognize the third, second and first nucleotides of the target sequence in the major groove of 5′→3′ DNA strand (Sera and Uranga, 2002). Thus, by joining 6 ZF domains together, 18 base pairs of DNA can be targeted, which are unique in the genome (Gersbach *et al.*, 2014).

Gallego-Bartolomé *et al.* (2018) targeted the *Arabidopsis FWA* promoter using fused human TET1cd to artificial ZFs along with control plants expressing a fusion of ZF108 to the fluorescent protein YPet (ZF108-YPet). Earlier demethylation of the promoter in *met1* mutants for fwa-4 epialleles is heritable across generations, triggers the ectopic expression of *FWA* and causes a late-flowering phenotype (Soppe *et al.*, 2000). Among the T_1 plants expressing ZF108-TET1cd in the Col-0 background, 25 of 57 displayed a late-flowering phenotype, suggesting *FWA* activation. Analyses of the flowering time of T_3 lines that either retained the ZF108-TET1cd transgene ($T3^+$) or had the transgene segregated away in the T_2 generation ($T3^-$) showed that both the $T3^+$ and $T3^-$ lines retained a delayed-flowering phenotype, which is consistent with hypomethylation at the *FWA* promoter. Expression of ZF108-TET1cd causes late flowering and *FWA* activation. However,

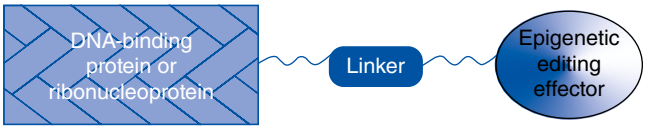


Fig. 5.7. Basic structure of an epigenome editing tool.

YPet (ZF108-YPet) did not show any effect on flowering time, suggesting that the delayed-flowering phenotype observed is not simply a consequence of ZF108 binding to the *FWA* promoter. FWA expression was dramatically increased in ZF108-TET1cd compared with Col-0 and ZF108-YPet and had a similar expression level to fwa-4, indicating that the late-flowering phenotype observed was attributable to *FWA* overexpression.

To test if the late-flowering phenotype observed was due to FWA upregulation, Gallego-Bartolomé *et al.* (2018) performed RNA-seq of Col-0, fwa-4 and four representative late-flowering T_1 plants expressing ZF108-TET1cd, as well as four biological replicates of Col-0, and two representative T3 lines expressing ZF108-TET1cd or ZF108-YPet. FWA expression was dramatically increased in ZF108–TET1cd compared with Col-0 and ZF108-YPet and had a similar expression level as *fwa-4*, indicating that the late-flowering phenotype observed was due to FWA overexpression. A genome-wide gene expression analysis showed very few additional changes and revealed *FWA* as the most upregulated gene in the ZF108-TET1cd lines compared with ZF108-YPet control lines. This study, thus suggested successful removal of methylation at the *FWA* promoter and, importantly, very few off-target effects attributable to ZF108-TET1cd expression.

Transcription activator-like effector-based epigenome editing tools

After ZFPs, TALEs are another class of DNA-binding proteins having three special structural features: first, they are an array of 33 or 34 repeated amino acid motifs; second, residues 12 and 13 as repeat variable di-residues (RVDs); and third, they have one truncated repeat with 20 amino acids. RVD from each monomer recognizes one nucleotide within the DNA-binding site (HD = C, NI = A, NG = T, NN = G). Similar to ZFPs, a TALE DBD can also be conjugated with an effector domain for targeted gene expression modification. The TALE DBD targeting efficiencies vary from 25 to 95% (Miller *et al.*, 2011; Maeder *et al.*, 2013). However, new assembly methods are also available to generate more efficient TALEs (Briggs *et al.*, 2012; Reyon *et al.*, 2012). Unlike ZFPs, TALEs allow single-base recognition of DNA rather than triplet-confined ZFPs; thus providing greater design flexibility. The design of TALE-DNA recognition is easy because specific di-residues for each nucleotide have been identified (Zhang *et al.*, 2011; Reyon *et al.*, 2012). In spite of target specificity and prior binding verification required in the case of TALE-DBD (Guilinger *et al.*, 2014), TALEs come next to ZFPs in gene expression modulation as proven by studies (Sanjana *et al.*, 2012; Gaj *et al.*, 2013).

Recently, Mlambo *et al.* (2018) described a novel platform, called a designer epigenome modifier (DEM), for achieving precise epigenome editing. DEMs combine in a single molecule as a DBD based on highly specific TALEs and several EDs capable of inducing DNA methylation and locally altering the chromatin structure to silence target gene expression. Mussolino (2018) has discussed the efficiency of DEMs and highlighted their remarkable safety profile.

CRISPR/dCas9-based epigenome editing tools

CRISPR are repeats (five 29-nt) interspersed with unique sequences (32-nt), first discovered in model bacteria, *Escherichia coli* (Ishino *et al.*, 1987). However, the function of these repeats was validated in *Streptococcus thermophilus*, providing adaptive immunity against infection by phages (Barrangou *et al.*, 2007; Deveau *et al.*, 2008). The CRISPR mechanism includes transcription of a repeat-spacer array into a precursor CRISPR RNA (pre-crRNA), which later gets processed into mature crRNAs. There are several pathways of CRISPR activation, one of which requires a trans-activating crRNA (tracrRNA). The pre-crRNA and tracrRNA base pair with each other to form an RNA duplex, which gets further cleaved by RNase III (ribonuclease) to form a final crRNA : tracrRNA hybrid structure. This hybrid RNA complex guides Cas9 endonuclease, which cleaves both the strands of the invading nucleic acid (Jinek *et al.*, 2012) and also determines Cas9 binding specificity. In this hybrid structure, 20 bases of crRNA are complementary to the respective target site and it requires protospacer-adjacent motif (PAM) sequence as a necessary component. One needs to only insert the desired DNA oligonucleotide into a vector construct for target site selection. In comparison to ZFPs and TALEs, the CRISPR/Cas9 technique is much easier, cheaper and less technical and more target-specific, as it shows Watson–Crick

pairing with target DNA. Moreover, simultaneous expression of multiple gRNAs allows multiplexing, which reduces the cost and time needed to generate plants with multiple targeted mutations.

To bring CRISPR/Cas9 into use for epigenome editing, the nuclease activity of Cas9 was abolished by creating two silencing mutations of the RuvC1 (D10A) and HNH (H841A) nuclease domains (Qi *et al.*, 2013); and this version of Cas9 was referred to as 'dead', 'deactivated' or 'nuclease-deficient' Cas9 (dCas9).

Targeted Epigenome Editing Procedures in Plants

Much basic research must be undertaken before opting for the use of epigenome editing technology for crop improvement. The scientist has to ensure that the trait to be improved is epigenetically inherited. Genome and epigenome maps of the plant species should be available. A particular epigenetic change(s) must be related with a phenotypic change. The target site in the genome has to be carefully selected to ensure the desired epigenetic change. Then comes the key decision of selecting, designing and constructing the epigenome editing tool. In plants, efficient epigenome editing generally comprises four continuous steps. First, the design and construction of a target-specific DBD (such as ZFP or TALE), or target-specific sgRNA in a CRISPR/dCas9 system. Many computer-based online tools have been developed to design DNA-binding domains and sgRNAs. However, the *in silico* design of the DBDs still needs further study for plants, and sgRNA efficiencies in plant cells are still needed to increase the accuracy of computational sgRNA selection. Second, the pre-validation of DBDs is required and best validated in protoplasts before being used for epigenome editing. Third, the components of the epigenome editing platforms are delivered into plant cells, normally via *Agrobacterium*-mediated transformation or particle bombardment, and the epigenome editing platform expression cassettes are stably integrated into the plant genome. Finally, transformed or regenerated plants with the desired modifications are identified by PCR genotyping and confirmed by sequencing.

DNA-free epigenome editing can be performed by delivering the ZFP- and TALE-based epigenome editing tools into the plant cell in protein or RNA form. In the case of the CRISPR/dCas9-based epigenome editing tool, the *in vitro* synthesized CRISPR/dCas9 protein sgRNA transcripts are pre-assembled in ribonucleoprotein (RNP) form, which can be delivered into immature embryos via particle bombardment. Alternatively, pre-assembled CRISPR/Cas9 RNP can be transfected into plant protoplasts. Bombarded/transfected cells are induced to form calli, from which seedlings are regenerated under the selection-free conditions. Regenerated plants are screened for desired epigenome changes. Delivering CRISPR/dCas9 reagents via RNP limits their temporal activity, thereby improving their precision.

Allele-specific Epigenome Editing

With such advance developments, allele-specific epigenome editing has accessed 'super-specific' variation, in which chromatin marks at only one selected allele of the target genomic locus can be altered (Bashtrykov and Jeltsch, 2018). This may prove to be a useful approach for the treatment of diseases caused by a mutant allele with a dominant effect, because silencing of the mutated allele would leave the healthy counterpart expressed. So, ultimately through this approach, direct correction of aberrant imprints in single alleles to correct imprinting disorders can be achieved.

General Applications of Epigenome Editing

Epigenome editing is a useful technique for altering the epigenome. Epigenome editing is now becoming the next substrate for achieving target traits. There are many applications of epigenome editing, including targeted gene upregulation or downregulation through editing of regulatory regions, such as promoters and enhancers (Fig. 5.8). Moreover, by modulating methylation across the promoter region, gene activation or repression can be achieved. Apart from this, epigenome editing can also be a useful tool to

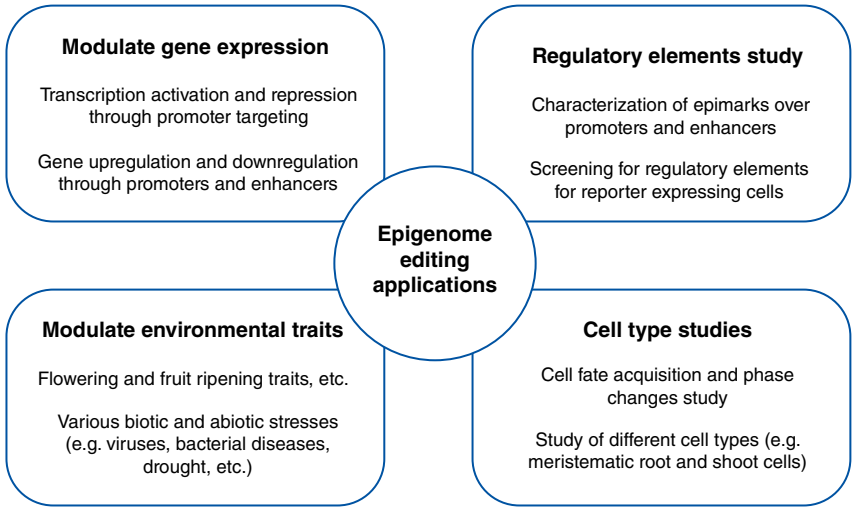


Fig. 5.8. Applications of epigenome editing.

study differentiation of meristematic cells, and root and shoot tissues in plants. As most of the agronomic traits in plants, such as flowering, ripening and biotic and abiotic stresses, are controlled by the environment, by modulating various epigenetic factors to play a role in such agronomic traits, we can produce the desired phenotype. With this epigenome editing technology, new sources of epigenetic variation can be created and various genes can be correctly regulated.

Limitations of Epigenome Editing Technology

Enríquez (2016) points out several limitations of epigenome editing technology, such as poor specificity of histone modifiers toward histone substrates, off-target effects, non-specific hyperacetylation of nucleosomal substrates and functions other than transcriptional repression. With respect to the claim that highly specific epigenome editing can be achieved across promoters and enhancers genome-wide, the data presented by the CRISPR/dCas9-p300 Core study merely corroborate targeting of desired DNA loci, not the deposition of unique epigenetic marks on histone tails. p300 HAT has poor specificity toward histone substrates. It is likely that other putative acetylation events would have been detected, had chromatin immunoprecipitation, followed by quantitative PCR (ChIP-qPCR), been performed using other histone PTM antibodies.

Despite evidence for robust transactivation at target genes using dCas9-p300 Core (Hilton *et al.*, 2015), it can hardly be said that highly specific epigenome editing takes place in the transfected cell lines. The touchstone principle of epigenome editing concerns uncovering specific functional roles of different PTMs. Besides transactivation of the target genes, non-H3K27 acetylation could likely have unspecified functional roles, including recruitment of specific readers of other acetylation marks, structural effects on higher order chromatin structure, or PTM-dependent signalling cascades. This highlights a key obstacle for current CRISPR-directed epigenome editing technologies. Contrary to the prevailing wisdom in genome editing, mitigation of off-target effects in epigenome editing is not exclusively related to the DNA-binding activity of dCas9 at specific loci, as the off-target effects have also been found to be related to deposition of PTM marks at specified nucleosomes with spatiotemporal precision. Some commercially available histone PTM antibodies cannot frequently detect their intended targets (Rothbart *et al.*, 2015). Although the issue of histone antibody specificity arguably applies to many epigenetic experiments, the problem is of particular importance in dCas9 fusions to catalytic domains.

The potential for non-specific hyperacetylation of nucleosomal substrates by dCas9-p300 Core also raises concerns about antibody specificity during epigenetic experiments. Based on the observations of Brocken *et al.* (2018), hyperacetylation of proximal nucleosomes by a tethered dCas9-p300 Core HAT might confound ChIP-qPCR experiments using anti-H3K27ac antibodies. Thus, potential non-H3K27 acetylation events may be playing a role in detecting highly enriched regions of H3K27ac at the enhancers and promoters surveyed. These principles applicable to transcriptional activation also apply to epigenome editing involving transcriptional repression and regulation of gene expression. Similar to HATs, other enzymes, including histone deacetylases, methyltransferases, demethylases, kinases, phosphatases, ubiquitin and SUMO ligases, have varying degrees of substrate specificity, and involve the use of commercial antibodies to detect PTM marks deposited on histones.

SUMOylation has, in general, been linked to transcriptional repression (Neyret-Kahn *et al.*, 2013; Decque *et al.*, 2016); however, its new functional roles, such as activation of DNA damage-signalling cascades, have also been reported (Wu *et al.*, 2014). Like ubiquitination, SUMOylation is likely to be involved in a widespread range of context-dependent, substrate functional roles, including proteasome-dependent proteolysis, signalling, assembly, cellular localization, and others (Deshaies and Joazeiro, 2009). Consequently, dCas9-KRAB-mediated epigenome editing aimed at repression can be confounded by physical and biochemical events that make the specific contributions of each component of the system's repressing machinery less clear. This may have serious repercussions, as CRISPR-dCas9-KRAB-mediated epigenome editing is capable of eliciting a desired silencing event, but triggers a deleterious secondary effect caused by unknown or unforeseen processes. KRAB's corepressor, KAP1, is known to recruit many factors and complexes associated with repressive epigenetic states. For example, heterochromatin protein 1 (HP1) is a protein complex which is vital for the formation of transcriptionally inactive heterochromatin (Ying *et al.*, 2015). SETDB1 is a methyltransferase that catalyses deposition of methyl histone marks on H3K9 (Schultz *et al.*, 2002). NuRD is a corepressor complex that catalyses histone deacetylation and ATP-dependent chromatin remodelling (Xue *et al.*, 1998) and N-COR1 – a histone deacetylase complex (Underhill *et al.*, 2000). KAP1 has also been linked to functional roles in DNA double-stranded repair (Noon *et al.*, 2010), restriction of retrovirus replication (Rowe *et al.*, 2010) and regulation of the self-renewal process in embryonic stem cells (Hu *et al.*, 2009). Given the vast range of KRAB- and KAP1-mediated interactions, scientists may foresee hurdles associated with epigenome editing, particularly when repression is orchestrated by multiple enzymes of varying degrees of substrate specificity, and proteins involved in long-range interactions. The biggest challenge in epigenome editing has been posed by the difficulty in packaging of dCas9 fusion proteins into the adeno-associated virus (AAV) used as the delivery vehicle (Lau and Suh, 2018).

Future Prospects of Epigenome Editing in Crop Improvement

Epigenome editing will help to harness useful epigenetic variation. Epigenetic variation observed in *ddm1* and *met1* mutants provided the basis for creation of epigenetic recombinant inbred lines (epiRILs). The epiRIL populations provide a unique opportunity to reduce genetic variability and increase epigenetic variability (Schmitz and Ecker, 2012). Epigenome editing will enable future crops to cope with adverse environmental conditions. Desired gene expression will be achieved through precise epigenome editing of their regulatory elements. With epimutations, the range of phenotypes can be large. Epigenetics can play a role in crop improvement for selection of favourable epigenetic states. Plants could temporarily activate or deactivate genes in response to change in the environment and thus could help in evolving varieties that could face the challenges of a changing environment. Depending on climatic conditions, genes could be switched on or off in plants. What if plants knew which genes to switch on and off in a certain climate? That is what was investigated by team of researchers at the Salk Institute for Biological Studies, in a series of studies on the model plant *A. thaliana* – a common mustard weed (Stecker, 2013). A deeper understanding

of this hidden layer of genetic diversity could bring further advances in plant breeding and bioengineering.

Groundbreaking achievements in the field of epigenetics are yielding insights into the role of each and every individual chromatin mark in the context of maintaining the cellular phenotype and regulating transient gene expression changes (Cano-Rodriguez and Rots, 2016). Thus, by knowing the role of individual chromatin marks, we are now on the verge of reconstructing gene expression at will.

Epigenome editing is an upcoming approach for inducing desired and durable gene regulation, with wider scope than conventional approaches used to determine regulatory functions and functions of chromatin modifications and cellular reprogramming (Kungulovski and Jeltsch, 2016). Epigenetic editing tools enable us to predict the effect of epigenetic modifications on gene expression. This information can be utilized to manipulate cell fate for both basic and applied research. The targeted rewriting and/or erasing of epigenetic modifications reconstruct local chromatin structure, with the potential to stimulate long-lasting changes in gene transcription. Each of these platforms has advantages and disadvantages with regard to genomic specificity, potency in regulating gene expression and reprogramming cell phenotypes, as well as ease of design, construction and delivery (Waryah *et al.*, 2018).

In addition to genomic diversity, epigenetic diversity provides additional sources of variation within a species that could be captured or created for crop improvement. We still need to understand such sources of epigenetic variation and their stability to improve crops (Springer and Schmitz, 2017). The recent developments in epigenome profiling and engineering may create new avenues for using the full potential of epigenetics in crop improvement. Endogenous regulatory elements existing in present crop species are not able to modulate their expression according to environmental changes, so epigenome editing may help in this regard.

In epigenome editing, presently, there are many considerations regarding specific design depending on the scientific question. Important design considerations and challenges regarding the biochemical- and locus-specificities of epigenome editors include how to: account for the complex biochemical diversity of chromatin; control for potential interdependency of epigenome editors and their resultant modifications; avoid sequestration effects; quantify the locus specificity of epigenome editors; and improve locus-specificity by considering concentration, affinity, avidity and sequestration effects (Sen and Keung, 2018).

The successful applications of epigenome-editing technology depend on various critical parameters like specificity, effectivity and sustainability of epigenome editing in experimental settings. In addition, conditions, such as the expression levels and the duration of the expression of the epi-editors, their DNA-binding affinity and specificity, and the cross talk between epi-editors and cellular chromatin modifiers (Rots and Jeltsch, 2018), are other factors that need to be considered for successful use of epigenome editing technology. Once established in a fully functional form, epigenome editing will allow better understanding of epigenetic expression control and translation of such knowledge into tools useful in crop improvement.

Understanding chromatin reprogramming underlying cell fate changes can pave the way for future crop manipulation. Epigenetic states in plants, once established during callus formation, can be inherited through the transmission of epigenetic alleles across generations (Lee and Seo, 2018). Epigenetic editing and engineering could be the key strategies for future crop manipulation. In-depth study of such epigenetic changes during cellular differentiation and cell fate acquisition in responsive crop species may also allow us to bring similar epigenetic marks in non-responsive crop species to break their recalcitrant character.

Recent researches have focused on the roles of miRNAs in epigenetic regulation of stress/adaptive responses as well as in providing plant genome stability (Xu *et al.*, 2018). Several potential stress responsive miRNAs are being studied from different crop plants and miRNA-driven RNA-interference (RNAi) is a choice for improving crop traits and providing phenotypic plasticity in challenging environments. Exploration of miRNAs as potent targets for engineering crops that can withstand multi-stress environments via loss-/gain-of-function approaches is in progress. The potential roles of plant miRNAs in genome stability and their emergence as potent target for genome editing are being discovered.

Experiments conducted so far on the use of epigenome editing in crop improvement have revealed encouraging results. Future work will reveal if epigenome editing fulfils its great promise in basic research and potentials of its usefulness in crop improvement. The CRISPR/Cas9 system has multiple benefits and that is why scientists mostly select this system for epigenome editing in several biological systems (Khan *et al*., 2017). Crop improvement can and will be greatly impacted by the use of techniques related to epigenetics, epigenomics and the newly emerging field of epibreeding (Kapazoglou *et al*., 2018). The successful use of induced epigenetic modifications in plants to improve the yield of field crops with minimal side effects will depend on how well we understand the connection between epigenetic change and phenotypic change.

References

Akimoto, K., Katakami, H., Kim, H.J., Ogawa, E., Sano, C.M., *et al*. (2007) Epigenetic inheritance in rice plants. *Annals of Botany* 100, 205–217.

Amaral, P.P., Dinger, M.E., Mercer, T.R. and Mattick, J.S. (2008) The eukaryotic genome as an RNA machine. *Science* 319, 1787–1789.

Angers, B., Castonguay, E. and Massicotte, R. (2010) Environmentally induced phenotypes and DNA methylation: How to deal with unpredictable conditions until the next generation and after. *Molecular Ecology* 19, 1283–1295.

Aufsatz, W., Mette, M.F., van der Winden, J., Matzke, A.J. and Matzke, M. (2002) RNA-directed DNA methylation in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* 99, 16499–16506.

Bai, S., Kasai, A., Yamada, K., Li, T. and Harada, T. (2011) A mobile signal transported over a long distance induces systemic transcriptional gene silencing in a grafted partner. *Journal of Experimental Botany* 62, 4561–4570.

Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Research* 21, 381–395.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., *et al*. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712.

Bashtrykov, P. and Jeltsch. A. (2018) Allele-specific epigenome editing. *Methods in Molecular Biology* 1767, 137–146.

Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes & Development* 16(1), 6–21.

Biswas, S. and Rao, C.M. (2018) Epigenetic tools (the writers, the readers and the erasers) and their implications in cancer therapy. *European Journal of Pharmacology* 837, 8–24.

Bond, D.M. and Baulcombe. D.C. (2014) Small RNAs and heritable epigenetic variation in plants. *Trends in Cell Biology* 24(2), 100–107.

Briggs, A.W., Rios, X., Chari, R., Yang, L., Zhang, F., *et al*. (2012) Iterative capped assembly: Rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic acids Research* 40, e117. DOI: 10.1093/nar/gks624.

Brocken, D.J.W., Tark-Dame, M. and Dame, R.T. (2018) dCas9: A versatile tool for epigenome editing. *Current Issues in Molecular Biology* 26, 15–32.

Cano-Rodriguez, D. and Rots, M.G. (2016) Epigenetic editing: On the verge of reprogramming gene expression at will. *Current Genetic Medicine Reports* 4(4), 170–179.

Cao, X. and Jacobsen, S.E. (2002) Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. *Proceedings of the National Academy of Sciences of the United States of America* 99, 16491–16498.

Chavez-Blanco, A., Perez-Plasencia, C., Perez-Cardenas, E., Carrasco-Legleu, C., Rangel-Lopez, E., *et al*. (2006) Antineoplastic effects of the DNA methylation inhibitor hydralazine and the histone deacetylase inhibitor valproic acid in cancer cell lines. *Cancer Cell International* 6, 2. DOI: 10.1186/1475-2867-6-2.

Chen, W., Kong, J., Lai. T., Manning. K., Wu, C., *et al*. (2015) Tuning LeSPL-CNR expression by Sly-miR157 affects tomato fruit ripening. *Science Reporter* 5, 7852. DOI: 10.1038/srep07852.

Chen, Y.T., Shen, C.H., Lin, W.D., Chu, H.A., Huang, B.L., *et al*. (2013) Small RNAs of *Sequoia sempervirens* during rejuvenation and phase change. *Plant Biology* 15(1), 27–36.

Chinnusamy, V. and Zhu, J.K. (2009) Epigenetic regulation of stress responses in plants. *Current Opinion in Plant Biology* 12, 133–139.

Choo, Y., Sanchez-Garcia, I. and Klug, A. (1994) *In vivo* repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* 372, 642–645.

Collins, L.J. and Chen, X.S. (2009) Ancestral RNA: The RNA biology of the eukaryotic ancestor. *RNA Biology* 6, 1–8.

Collins, L.J. and Penny, D. (2009) The RNA infrastructure: Dark matter of the eukaryotic cell? *Trends in Genetics* 25, 120–128.

Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., *et al*. (2014) Mapping the epigenetic basis of complex traits. *Science* 343, 1145–1148.

Decque, A., Joffer, O., Magalhaes, J.G., Cossec, J.C., Blecher-Gonen, R., *et al*. (2016) Sumoylation coordinates the repression of inflammatory and anti-viral gene-expression programmes during innate sensing. *Nature Immunology* 17, 140–149.

De-la-Peña, C., Nic-Can, G., Ojeda, G., Herrera-Herrera, J.L., López-Torres, A., *et al*. (2012) *KNOX1* is expressed and epigenetically regulated during *in vitro* conditions in *Agave* spp. *BMC Plant Biology* 12, 203. DOI: 10.1186/1471-2229-12-203.

Deshaies, R.J. and Joazeiro, C.A. (2009) RING domain E3 ubiquitin ligases. *Annual Review of Biochemistry* 78, 399–434.

Deveau, H., Barrangou, R., Garneau. J.E., Labonte. J., Fremaux. C., *et al*. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology* 190, 1390–1400.

Eichten, S.R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., *et al*. (2013) Epigenetic and genetic influences on DNA methylation variation in maize populations. *The Plant Cell* 25, 2783–2797.

Enríquez, P. (2016) CRISPR-mediated epigenome editing. *The Yale Journal of Biology and Medicine* 89(4), 471–486.

Espinas, N.A., Saze, H. and Saijo, Y. (2016) Epigenetic control of defense signaling and priming in plants. *Frontiers in Plant Science* 7, 1201. DOI: 10.3389/fpls.2016.01201.

Feng, Q., Yang, C., Lin, X., Wang, J., Ou, X., *et al*. (2012) Salt and alkaline stress induced transgenerational alteration in DNA methylation of rice (*Oryza sativa*). *Australian Journal of Crop Science* 6, 877–883.

Feng, S., Jacobsen, S.E. and Reik, W. (2010) Epigenetic reprogramming in plant and animal development. *Science* 330(6004), 622–627.

Gaj, T., Mercer, A.C., Sirk, S.J., Smith, H.L. and Barbas, C.F. III. (2013) A comprehensive approach to zinc-finger recombinase customisation enables genomic targeting in human cells. *Nucleic Acids Research* 41, 3937–3946.

Gallego-Bartolomé, J.G., Gardiner, J., Liu, W., Papikian, A., Ghoshal, B., *et al*. (2018) Targeted DNA demethylation of the *Arabidopsis* genome using the human TET1 catalytic domain. *Proceedings of the National Academy of Sciences of the United States of America* 115(9), E2125–E2134.

Gallusci, P., Dai, Z., Génard, M., Gauffretau, A., Fournier, N.L., *et al*. (2017) Epigenetics for plant improvement: Current knowledge and modeling avenues. *Trends in Plant Science* 22, 610–623.

Gersbach, C.A., Gaj, T. and Barbas, C.F. (2014) Synthetic zinc finger proteins: The advent of targeted gene regulation and genome modification technologies. *Accounts of Chemical Research* 47, 2309–2318.

Guilinger, J.P., Thompson, D.B. and Liu, D.R. (2014) Fusion of catalytically inactive Cas9 to *Fok*I nuclease improves the specificity of genome modification. *Nature Biotechnology* 32, 577–582.

Hardcastle, T.J., Müller, S.Y. and Baulcombe, D.C. (2018) Towards annotating the plant epigenome: The *Arabidopsis thaliana* small RNA locus map. *Scientific Reports* 8, 6338. DOI: 10.1038/s41598-018-24515-8.

Hauben, M., Haesendonckx, B., Standaert, E., Van Der Kelen, K., Azmi, A., *et al*. (2009) Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. *Proceedings of the National Academy of Sciences of the United States of America* 106(47), 20109–20114.

Hauser, M.T., Aufsatz, W., Jonak, C. and Luschnig, C. (2011) Transgenerational epigenetic inheritance in plants. *Biochimica et Biophysica Acta* 1809, 459–468.

Havecker, E.R., Wallbridge, L.M., Hardcastle, T.J., Bush, M.S., Kelly, K.A., *et al*. (2010) The *Arabidopsis* RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. *The Plant Cell* 22, 321–334.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., *et al*. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333(6047), 1303–1307.

Heiderscheit, E.A., Eguchi, A., Spurgat, M.C. and Ansari, A.Z. (2018) Reprogramming cell fate with artificial transcription factors. *FEBS Letters* 592(6), 888–900.

Henderson, I.R. and Jacobsen, S.E. (2007) Epigenetic inheritance in plants. *Nature* 447, 418–424.

Heo, J.B., Lee, Y. and Sung, S. (2013) Epigenetic regulation by long noncoding RNAs in plants. *Chromosome Research* 21, 685–693.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., *et al*. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology* 33, 510–517.

Holliday, R. and Pugh, J.E. (1975) DNA modification mechanisms and gene activity during development. *Science* 187, 226–232.

Hu, G., Kim, J., Xu, Q., Leng, Y., Orkin, S.H., *et al*. (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes & Development* 23, 837–848.

Hu, L., Li, N., Xu, C., Zhong, S., Lin, X., *et al*. (2014) Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proceedings of the National Academy of Sciences of the United States of America* 111, 10642–10647.

Huanca-Mamani, W., Arias-Carrasco, R., Cárdenas-Ninasivincha, S., Rojas-Herrera, M., Sepúlveda-Hermosilla, G., *et al*. (2018) Long non-coding RNAs responsive to salt and boron stress in the hyperarid Lluteño maize from Atacama Desert. *Genes* 9(3), 170. DOI: 10.3390/genes9030170.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology* 169, 5429–5433.

Iwasaki, M. and Paszkowski, J. (2014) Identification of genes preventing transgenerational transmission of stress-induced epigenetic states. *Proceedings of the National Academy of Sciences of the United States of America* 111(23), 8547–8552.

Ji, L., Neumann, D.A. and Schmitz, R.J. (2015) Crop epigenomics: Identifying, unlocking, and harnessing cryptic variation in crop genomes. *Molecular Plant* 8, 860–870.

Ji, L., Jordan, W.T., Shi, X., Hu, L., He, C., *et al*. (2018) TET-mediated epimutagenesis of the *Arabidopsis thaliana* methylome. *Nature Communications* 9, 895. DOI: 10.1038/s41467-018-03289-7.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., *et al*. (2012) A programmemable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821.

Johannes, F., Porcher, E., Teixeira, F.K., Saliba–Colombani, M., Simon, V., *et al*. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics* 5, e1000530. DOI: 10.1371/journal.pgen.1000530.

Jones, L., Ratcliff, F., and Baulcombe, D.C. (2001) RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. *Current Biology* 11, 747–757.

Jurkowska, R.Z., Jurkowski, T.P. and Jeltsch, A. (2011) Structure and function of mammalian DNA methyltransferases. *Chembiochem* 12, 206–222.

Kanherkar, R.R., Bhatia-Dey, N. and Csoka, A.B. (2014) Epigenetics across the human life span. *Frontiers in Cell and Developmental Biol*ogy 2, 1–19.

Kapazoglou, A., Ganopoulos, I., Tani, E. and Tsaftaris, A. (2018) Epigenetics, epigenomics and crop improvement. *Advances in Botanical Research* 86, 287–324.

Kasai, A., Bai, S., Hojo, H. and Harda, T. (2016) Epigenome editing of potato by grafting using transgenic tobacco as siRNA Donor. *PLoS ONE* 11, 1–12.

Kasai, M. and Kanazawa, A. (2013) Induction of RNA-directed DNA methylation and heritable transcriptional gene silencing as a tool to engineer novel traits in plants. *Plant Biotechnology* 30, 233–241.

Khan, H.M.U., Khan, S.U., Muhammad, A., Hu, L., Yang, Y., *et al*. (2017) Induced mutation and epigenetics modification in plants for crop improvement by targeting CRISPR/Cas9 technology. *Journal of Cellular Physiology* 233(6), 4578–4594.

Köferle, A., Stricker, S.H. and Beck, S. (2015) Brave new epigenomes: The dawn of epigenetic engineering. *Genome Medicine* 7(1), 59. DOI:10.1186/s13073-015-0185-8.

Kong, L., Zhang, Zhi-Qiang, Y., Liu, X., Zhao, S., and Wei, L. and Gao, G. (2007) CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* 35(Suppl. 2), W345–W349.

Kouzarides, T. (2007) Chromatin modifications and their function. *Cell* 128, 693–705.

Ku, Y.S., Wong, J.W., Mui, Z., Liu, X., Hui, J.H., Chan, T. and Lam, H. (2015) Small RNAs in plant responses to abiotic stresses: Regulatory roles and study methods. *International Journal of Molecular Sciences* 16, 24532–24554.

Kungulovski, G. and Jeltsch, A. (2016) Epigenome editing: State of the art, concepts, and perspectives. *Trends in Genetics* 32(2), P101–P113.

Kungulovski, G., Nunna, S., Thomas, M., Zanger, U.M., Reinhardt, R. and Jeltsch, A. (2015) Targeted epigenome editing of an endogenous locus with chromatin modifiers is not stably maintained. *Epigenetics and Chromatin* 8, 12. DOI: 10.1186/s13072-015-0002-z.

Lau, C. and Suh, Y. (2018) *In vivo* epigenome editing and transcriptional modulation using CRISPR technology. *Transgenic Research* 27(6), 489–509.

Law, J.A., Ausin, I., Johnson, L.M., Vashisht, A.A., Zhu, J.K., *et al*. (2010) A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in *Arabidopsis*. *Current Biology* 51, 951–956.

Lee, K. and Seo, P.J. (2018) Dynamic epigenetic changes during plant regeneration. *Trends in Plant Science* 23(3), 235–247.

Levine, R. (2019). Epigenetics could alter the way we breed crops for drought and climate change. Available at: https://geneticliteracyproject.org/2019/04/19/epigenetics-could-alter-the-way-we-breed-crops-for-drought-and-climate-change/?fbclid=IwAR1BPqgXoQhAV1gRU03O2CYJROEtpCh5ShyAswskYaUt5uuLCB4b34YaZek (accessed 25 April 2019).

Li, L., Wang, X., Sasidharan, R., Stolc, V., Deng, W., *et al*. (2007) Global identification and characterization of transcriptionally active regions in the rice genome. *PLoS ONE* 2(3), e294. DOI: 10.1371/journal.pone.0000294.

Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C., *et al*. (2014a) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology* 15, R40. DOI: 10.1186/gb-2014-15-2-r40.

Li, M, Liu, Y., Zhang, X., Liu, J. and Wang, P. (2018a) Transcriptomic analysis of high-throughput sequencing about circRNA, lncRNA and mRNA in bladder cancer. *Gene* 677, 189–197.

Li, Q., Eichten, S.R., Hermanson, P.J., Zaunbrecher, V.M., Song, J., *et al*. (2014b) Genetic perturbation of the maize methylome. *The Plant Cell* 26, 4602–4616.

Li, Q., Song, J., West, P.T., Zynda, G., Eichten, S.R., *et al*. (2015) Examining the causes and consequences of context-specific differential DNA methylation in maize. *Plant Physiology* 168, 1262–1274.

Li, R., Fu, D., Zhu, B., Luo, Y. and Zhu, H. (2018b) CRISPR/Cas9-mediated mutagenesis of lncRNA1459 alters tomato fruit ripening. *The Plant Journal* 94, 513–524.

Li, X., Xing, X., Xu, S., Zhang, M., Wang, Y., *et al*. (2018c) Genome-wide identification and functional prediction of tobacco lncRNAs responsive to root-knot nematode stress. *PLoS ONE* 13(11), e0204506. DOI: 10.1371/journal.pone.0204506.

Liang, G., He, H. and Yu D. (2012) Identification of nitrogen starvation-responsive miRNAs in *Arabidopsis thaliana*. *PLoS ONE* 7(11), e48951. DOI: 10.1371/journal.pone.0048951.

Lindroth, A.M., Cao, X. and Jackson, J.P. (2001) Requirement of chromomethylase3 for maintenance of CpXpG methylation. *Science* 292, 2077–2080.

Lister, R., Malley, R.C. and Tonti-Filippini, J. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536.

Maeder, M.L., Linder, S.J., Reyon, D., Angstman, J.F. and Fu, Y. (2013) Robust, synergistic regulation of human gene expression using TALE activators. *Nature Methods* 10(3), 243–245.

Martins, M., Sarmento, D. and M.M. Oliveira. (2004) Genetic stability of micropropagated almond plantlets, as assessed by RAPD and ISSR markers. *Plant Cell Reports* 23(7), 492–496.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B. and Matzke, A.J. (2009) RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology* 21, 367–376.

Matzke, M.A., Kanno, T. and Matzke, A.J.M. (2015) RNA-directed DNA methylation: The evolution of a complex epigenetic pathway in flowing plants. *Annual Review of Plant Biology* 66, 243–267.

McGarry, R.C. and Kragler, F. (2013) Phloem-mobile signals affecting flowers: Applications for crop breeding. *Trends in Plant Science* 18, 198–206.

Miglani, G.S. (2017) Genome editing in crop improvement: Present scenario and future prospectus. *Journal of Crop Improvement* 31(4), 453–559.

Miglani, G.S. (2019) *Genome editing: A Comprehensive Treatise*. Alpha Science International Limited, Oxford, UK.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., *et al*. (2011) A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology* 29, 143–148.

Mirouze, M. and Paszkowski, J. (2011) Epigenetic contribution to stress adaptation in plants. *Current Opinion in Plant Biology* 14, 267–274.

Mlambo, T., Nitsch, S. and Hildenbeutal, M. (2018) Designer epigenome modifiers enable robust and sustained gene silencing in clinically relevant human cells. *Nucleic Acids Research* 46, 4456–4468.

Molnar, A., Melnyk, C.W., Bassett, A., Hardcastle, T.J., Dunn, R., *et al*. (2010) Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *Science* 328, 872–875.

Mussolino, C. (2018) Precise epigenome editing on the stage: A novel approach to modulate gene expression. *Epigenetics Insights* 11, 2516865718818838. DOI: 10.1177/2516865718818838.

Nathan, M. and Robert, J. (2017) Exploiting induced and natural epigenetic variation for crop improvement. *Nature Reviews Genetics* 18, 563–575.

Neyret-Kahn, H., Benhamed, M., Ye, T., Le Gras, S., Cossec, J.C., *et al*. (2013) Sumoylation at chromatin governs coordinated repression of a transcriptional programme essential for cell growth and proliferation. *Genome Research* 23, 1563–1579.

Noon, A.T., Shibata, A., Rief, N., Löbrich, M., Stewart, G.S., *et al*. (2010) 53BP1-dependent robust localized KAP-1 phosphorylation is essential for heterochromatic DNA double-strand break repair. *Nature Cell Biology* 12(2), 177–184.

Ong-Abdullah, M., Ordway, J.M., Jiang, N., Ooi, S.E., Kok, S.Y., *et al*. (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 525, 533–537.

Parra, A.L., Yhebra, R.S., Sardinas, I.G. and Buela, L.I. (2001) Comparative study of the assay of *Artemia salina* L. and the estimate of the medium lethal dose (LD50 value) in mice, to determine oral acute toxicity of plant extracts. *Phytomedicine* 8(5), 395–400.

Pontes, O., Li, C.F. and Nunes, P.C. (2006) The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* 126, 79–92.

Puchta, H. (2005) The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. *Journal of Experimental Botany* 56, 1–14.

Qi, L.S., Larson M.H., Gilbert L.A., Doudna J.A., Weissman J.S., *et al*. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173–1183.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., *et al*. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nature Biotechnology* 30, 460–465.

Romanowska, J. and Joshi, A. (2019) From genotype to phenotype: Through chromatin. *Genes* 10, 76. DOI:10.3390/genes10020076.

Rothbart, S.B., Dickson, B.M., Raab, J.R., Grzybowski, A.T., Krajewski, K., *et al*. (2015) An interactive database for the assessment of histone antibody specificity. *Molecular Cell* 59(3), 502–511.

Rots, M.G. and Jeltsch, A. (2018) Editing the epigenome: Overview, open questions, and directions of future development. *Methods in Molecular Biology* 1767, 3–18.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., *et al*. (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* 463, 237–240.

Sahu, P.P., Sharma, N., Puranik S., and Prasad, M. (2013) Post-transcriptional and epigenetic arms of RNA silencing: A defense machinery of naturally tolerant tomato plant against Tomato Leaf Curl New Delhi Virus. *Plant Molecular Biology Reports* 32, 1151–1159.

Sanjana, N.E., Cong, L., Zhou, Y., Cunnif, M.M., Feng, G., *et al*. (2012) A transcription activator-like effector toolbox for genome engineering. *Nature Protocols* 7, 171–192.

Saraswat, S., Yadav, A.K., Sirohi, P. and Singh, N.K. (2017) Role of epigenetics in crop improvement: Water and heat stress. *Journal of Plant Biology* 60, 231–240.

Schmitz, R.J. and J.R. Ecker. (2012) Epigenetic and epigenomic variation in *Arabidopsis thaliana*. *Trends in Plant Science* 17(3), 149–154.

Schmitz, R.J., He, Y., Valdés-López, O., Khan, S.M., Joshi, T., *et al*. (2013a) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research* 23, 1663–1674.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., *et al*. (2013b) Patterns of population epigenomic diversity. *Nature* 495, 193–198.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G. and Rauscher, F.J., III. (2002). SETDB1: A novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc finger proteins. *Genes & Development* 16, 919–932.

Sen, D. and Keung, A.J. (2018) Designing epigenome editors: Considerations of biochemical and locus specificities. *Methods in Molecular Biology* 1767, 65–87.

Sera, T. and Uranga, C. (2002) Rational design of artificial zinc-finger proteins using a nondegenerate recognition code table. *Biochemistry* 41, 7074–7081.

Siddiqui, Z.H., Abbas, Z.K., Ansari, M.W. and Khan, M.N. (2018) The role of miRNA in somatic embryogenesis. *Genomics* 111, 1026–1033. DOI: 10.1016/j.ygeno.2018.11.022.

Singh, U., Khemka, N., Rajkumar, M.S., Garg, R. and Jain, M. (2017) PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Research* 45(22), e183. DOI: 10.1093/nar/gkx866.

Snowden, A.W., Gregory, P.D., Case, C.C. and Pabo, C.O. (2002) Gene-specific targeting of H3K9 methylation is sufficient for initiating repression *in vivo*. *Current Biology* 12(24), 2159–2166.

Soppe, W.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, *et al*. (2000) The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Molecular Cell* 6(4), 791–802.

Springer, N.M. and R.J. Schmitz. (2017) Exploiting induced and natural epigenetic variation for crop improvement. *Nature Reviews Genetics* 18, 563–575.

Stecker, T. (2013) Can epigenetics help crops adapt to climate change? *ClimateWire*, March 7. Available at: https://www.scientificamerican.com/article/can-epigentics-help-crops-adapt-to-climate-change (accessed 7 June 2019).

Stelpflug, S.C., Eichten, S.R., Hermanson, P.J., Springer, N.M. and Kaeppler, S.M. (2014) Consistent and heritable alterations of DNA methylation are induced by tissue culture in maize. *Genetics* 198(1), 209–218.

Stroud, H., Ding, B., Simon, S.A., Feng, S., Bellizzi, M., *et al*. (2013) Plants regenerated from tissue culture contain stable epigenome changes in rice. *eLife* 2, e00354. DOI: 10.7554/eLife.00354.

Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F., *et al*. (2018) METHimpute: Imputation-guided construction of complete methylomes from WGBS data. *BMC Genomics* 19, 444. DOI: 10.1186/s12864-018-4641-x.

Thakore, P.I., Black, J.B., Hilton, I.B. and Gersbach, C.A. (2016) Editing the epigenome: Technologies for programmable transcription and epigenetic modulation. *Nature Methods* 13(2), 127–137.

Underhill, C., Qutob, M.S., Yee, S.P. and Torchia, J. (2000) A novel nuclear receptor corepressor complex, N-CoR, contains components of the mammalian SWI/SNF complex and the corepressor KAP-1. *Journal of Biological Chemistry* 275, 40463–40470.

Us-Camas, R., Rivera-Solís, G., Duarte-Ake, F. and De-la-Peña, C. (2014) *In vitro* culture: An epigenetic challenge for plants. *Plant Cell, Tissue and Organ Culture* 118(2), 187–201.

Vanyushin, B.F. and Ashapkin, V.V. (2011) DNA methylation in higher plants: Past, present and future. *Biochimica et Biophysica Acta* 1809, 360–368.

Waddington, C.H. (1942) The epigenotype. *Endeavour* 1, 18–20.

Wang, H., Chung, P.J., Liu, J., Jang, I., Kean, M.J., *et al*. (2014) Genome-wide identification of long non-coding natural antisense transcripts and their responses to light in *Arabidopsis*. *Cytogenetic and Genome Research* 24, 444–453.

Wang, J., Meng, X., Dobrovolskaya, O.B., Orlov, Y.L. and Chen, M. (2017) Non-coding RNAs and their roles in stress response in plants. *Genomics, Proteomics and Bioinformatics* 15(5), 301–312.

Wang, K., Zhang, S.L., Weber, J., Baxter, D. and Galas, D.J. (2010) Export of microRNAs and microRNA-protective protein by mammalian cells. *Nucleic Acids Research* 38, 7248–7259.

Wang, Y., Luo, X., Sun, F., Hu, J., Zha, X., *et al*. (2018) Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. *Nature Communications* 9, 3516. DOI: 10.1038/s41467-018-05829-7.

Waryah, C.B., Moses, C., Arooj, M. and Blancafort, P. (2018) Zinc fingers, TALEs, and CRISPR systems: A comparison of tools for epigenome editing. *Methods in Molecular Biology* 1767, 19–63.

Weber, M. and Schubeler, D. (2007) Genomic patterns of DNA methylation: Targets and function of an epigenetic mark. *Current Opinion in Cell Biology* 19, 273–280.

Wierzbicki, A.T., Haag, J.R. and Pikaard, C.S. (2008) Non-coding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* 135, 635–648.

Wierzbicki, A.T., Ream, T.S., Haag, J.R. and Pikaard, C.S. (2009) RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature Genetics* 41, 630–634.

Wu, C.-S., Ouyang, J., Mori, E., Nguyen, H.D., Maréchal, A., *et al*. (2014) SUMOylation of ATRIP potentiates DNA damage signaling by boosting multiple protein interactions in the ATR pathway. *Genes & Development* 28, 1472–1484.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., *et al*. (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biology* 2(5), e104. DOI: 10.1371/journal.pbio.0020104.

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z., *et al*. (2011) Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biology* 11, 61. DOI: 10.1186/1471-2229-11-61.

Xu, C., Tian, J. and Mo, B. (2013) siRNA-mediated DNA methylation and H3K9 dimethylation in plants. *Protein Cell* 4, 656–663.

Xu, J., Hou, Q., Khare, T., Verma, S.K. and Kumar, V. (2018) Exploring miRNAs for developing climate-resilient crops: A perspective review. *Science of the Total Environment* 653, 91–104.

Xue, Y., Wong, J., Moreno, G.T., Young, M.K., Côté, J., *et al*. (1998) NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Molecular Cell* 2, 851–861.

Yang, B., Borgeaud, A., Aeschbach, L. and Dion, V. (2018) Uncovering the interplay between epigenome editing efficiency and sequence context using a novel inducible targeting system. *bioRxiv*. DOI: 10.1101/368480.

Ying, Y., Yang, X., Zhao, K., Mao, J., Kuang, Y., *et al*. (2015) The Krüppel-associated box repressor domain induces reversible and irreversible regulation of endogenous mouse genes by mediating different chromatin states. *Nucleic Acids Research* 43, 1549–1561.

Zhang, F., Cong, L., Lodato, S., Kosuri, S., Church, G.M., *et al*. (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature Biotechnology* 29, 149–153.

Zhang, L., Wang, Y., Zhang, X., Zhang, M., Han, D., *et al*. (2012) Dynamics of phytohormone and DNA methylation patterns changes during dormancy induction in strawberry (Fragaria × ananassa Duch.). *Plant Cell Reports* 31(1), 155–165.

Zhu, B., Yang, Y., Li, R., Fu, D., Wen, L., *et al*. (2015) RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. *Journal of Experimental Botany* 66, 4483–4495.

Zhu, Q.H. and Wang, M.B. (2012) Molecular functions of long non-coding RNAs in plants. *Genes (Basel)* 3, 176–190.

Zhu, Z., Hughes, K.W. and Huang, L. (1991) Effects of 5-azacytidine on transformation and gene expression in *Nicotiana tabacum*. *In Vitro Cellular & Developmental Biology – Plant* 27, 77–83.

6 Bioinformatics and Plant Breeding

Robyn Anderson, Cassandria Tay Fernandez, Monica F. Danilevicz and David Edwards*
The University of Western Australia, Perth, Australia

Introduction

With the fast-expanding global population and the changing climate, there is an urgent requirement to accelerate the production of high-yielding, climate-resilient varieties. The application of genetics and genomics, supported by advances in bioinformatics, offers approaches to accelerate breeding to produce advanced crops. Advances have been made through the discovery and application of genetic markers, the use of whole-genome sequence data, the association of agronomic traits with genomic variation and the implementation of this knowledge in applied breeding. Future potential, through advances in genome editing, supported by artificial intelligence and machine learning applications to agricultural research, is opening the way for a new era of accelerated plant breeding and custom crops. In this chapter, we detail how bioinformatics can be applied to support plant breeding to produce better crops faster, supporting global food security.

Genetics and Molecular Markers

The growth of bioinformatics has mirrored the growth in biological data, particularly the remarkable expansion of DNA-sequencing technology. These data have been applied for breeding from the early forms of molecular markers, Restriction fragment length polymorphisms (RFLPs) and isozymes; and some of the early computational tools were developed to produce genetic maps from these kinds of molecular markers, including JoinMap (Van Ooijen and Voorrips, 2001) and Mapmaker (Lander *et al.*, 1987). Genetic marker discovery was predominantly lab-based until the expansion of high-throughput, capillary-based Sanger sequencing machines, such as the ABI3730. The high throughput of these machines supported the establishment of large-scale sequencing of cDNAs in the form of expressed sequence tags (ESTs); for example, in the developing wheat grain (Wilson *et al.*, 2004), and the opportunity for high-throughput computational molecular-marker discovery.

Simple sequence repeats (SSRs or microsatellites) as well as single nucleotide polymorphisms (SNPs) are two of the earliest high-throughput sequence-based molecular markers, and tools were rapidly developed to mine the growing quantity of Sanger sequence data for these markers (Batley *et al.*, 2007b). SNP discovery software included autoSNP (Barker *et al.*, 2003; Batley *et al.*, 2003), SNPdetector (Zhang *et al.*, 2005), novoSNP (Weckx *et al.*, 2005) and the real-time SNP discovery tool SNPServer (Savage *et al.*,

* E-mail: dave.edwards@uwa.edu.au

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

2005a). This high-throughput Sanger sequence was also mined for SSR-based molecular markers using tools such as SSRPrimer (Robinson *et al.*, 2004), which was applied for the discovery of SSRs in both major and minor crop species, including maize, Brassicas and strawberry (Keniry *et al.*, 2006; Burgess *et al.*, 2006; Savage *et al.*, 2005b; Batley *et al.*, 2007a) and SSRPoly, which predicts polymorphic SSRs from sequence data (Duran *et al.*, 2013). With the continued growth of sequencing data and advanced SNP genotyping methods, the use of SSRs has now predominantly been replaced by SNPs, though with some important SSR loci being translated into SNP loci (Mogg *et al.*, 2002). The advent of next-generation DNA sequencing (NGS) opened up new opportunities for SNP marker discovery (Imelfort *et al.*, 2009; Edwards *et al.*, 2013) and several custom tools have been developed, including SGSautoSNP (Lai *et al.*, 2012a; Lorenc *et al.*, 2012).

The growth of molecular markers and the resulting genotype data led to the requirement for databases hosting this information and making it readily available to breeders and researchers (Batley and Edwards, 2009; Duran *et al.*, 2009c; Edwards *et al.*, 2014; Ruperao and Edwards, 2015; Lai *et al.*, 2015). Early formats included the crop ASTRA databases (Love *et al.*, 2005; Shields *et al.*, 2005; Spangenberg *et al.*, 2005a; Spangenberg *et al.*, 2005b), SSR Primer and SSR Taxonomy Tree (Jewell *et al.*, 2006), which hosted SSR marker data mined from the whole of Genbank, autoSNPdb (Duran *et al.*, 2009a, 2009b), and CerealsDB (Wilkinson *et al.*, 2016), which has gradually evolved with the growth and diversity of information. Other current molecular marker databases include CropSNPdb (Scheben *et al.*, 2019), which hosts genotype data from Illumina Infinium wheat and *Brassica* SNP arrays; Panzea (Zhao *et al.*, 2006), which hosts genotype data for wheat; Brassica.info and brassicagenome.net, which host data for *Brassica* species; CicArVarDB for chickpea (Doddamani *et al.*, 2015); Oryzabase for rice (Kurata and Yamazaki, 2006); Pea marker Database (Kulaeva *et al.*, 2017) for peas; and Sol Genomics Network for Solanaceae (Fernandez-Pozo *et al.*, 2015). There are also integrated databases, such as Gramene (Ware *et al.*, 2002a, 2002b; Jaiswal *et al.*, 2006; Youens-Clark *et al.*, 2011) and Graingenes (Matthews *et al.*, 2003; Carollo *et al.*, 2005; O'Sullivan, 2007), which host comprehensive information on cereal crops. More general databases, such as EnsemblePlants (Bolser *et al.*, 2016), PGDBj DNA Marker and Linkage Database (Asamizu *et al.*, 2014), PlantGDB (Dong *et al.*, 2004), Phytozome (Goodstein *et al.*, 2012) and GenBank (Benson *et al.*, 2009) exist. A summary of molecular databases and tools is listed in Table 6.1.

Genomics and Pan-genomics

Advancements in next-generation DNA sequencing moved studies from genetic to genomic analysis. Initial NGS data were still relatively expensive, especially where deep coverage was required for large crop genomes, and so methods to apply smaller sets of data were developed. TAGdb is a database established to permit the mining of short paired sequence reads based on identity to a query sequence (Marshall *et al.*, 2010), supporting the identification of the genomic sequence surrounding ESTs or polymorphisms between varieties. As data volumes increased, the genomes of an increasing number of crop species were assembled using advanced assembly algorithms and high-performance computing infrastructure. These included major crops, such as the Brassicas (Consortium, 2011; Chalhoub *et al.*, 2014; Liu *et al.*, 2014), as well as orphan crops, such as chickpea (Varshney *et al.*, 2013), lupin (Hane *et al.*, 2017) and clover (Kaur *et al.*, 2017). Even crops with very large and complex genomes could be approached using customized assembly algorithms and chromosome arm sequencing approaches. This was first demonstrated for the wheat group 7 chromosomes (Berkman *et al.*, 2011, 2012, 2013) and later to assemble the complete bread wheat genome (IWGSC, 2014) prior to assembly of the same genome in a single step (Appels *et al.*, 2018). While a combination of NGS and advanced bioinformatics tools enabled the assembly of an increasing number of crop genomes, the quality was often questionable. As a result, various bioinformatics approaches were developed to assess and validate these assemblies, including isolated chromosome sequencing (Ruperao *et al.*, 2014), optical mapping and associated software (Yuan *et al.*, 2017a, 2017b, 2018), and the skim-based genotyping by sequencing of populations (Bayer *et al.*, 2015; Golicz *et al.*, 2015b).

Table 6.1. A list of molecular-marker tools and databases.

Database name	Description	URl link	Reference
ASTRA databases	A database with the alternative splicing and alternative transcriptional initiation in humans, mice, flies, *Caenorhabditis elegans*, *Arabidopsis* and rice	Retired	(Love *et al*., 2005)
autoSNPdb	A SNP and EST sequence database for barley, *Brassica*, rice and wheat	http://autosnpdb.appliedbioinformatics.com.au/	(Duran *et al*., 2009a; Duran *et al*., 2009b)
Brassica.info	RFLP, SSR, Incel, SNP/InDel, AFLP, RAPD markers for *Brassica*	http://brassica.info/tools/genetic_markers.html	
brassicagenome.net	Database with pan-genome, SNPs and PAVs for *Brassica*	http://www.brassicagenome.net/databases.php	
CerealsDB	DArt markers and SNPs for the wheat genome	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php	(Wilkinson *et al*., 2016)
CicArVarDB	A chickpea SNP-InDel Database	https://cegresources.icrisat.org/cicarvardb/	(Doddamani *et al*., 2015)
CropSNPdb	SNP database for crop variation and *Brassica* and *Triticum aestivum*	http://snpdb.appliedbioinformatics.com.au/	(Scheben *et al*., 2019)
EnsemblPlants	A plant EST database with 50 species and 5000 EST markers	http://plants.ensembl.org/index.html	(Bolser *et al*., 2016)
GenBank	General database with ESTs and SNPs	https://www.ncbi.nlm.nih.gov/genbank/	(Benson *et al*., 2009)
GrainGenes	A database for Triticeae and *Avena*	https://wheat.pw.usda.gov/cgi-bin/GG3/browse.cgi?class=marker	(O'Sullivan, 2007)
Gramene	Database for crops and model plant species	http://www.gramene.org/	(Ware *et al*., 2002a; Ware *et al*., 2002b; Jaiswal *et al*., 2006)
Oryzabase	Rice database with a variety of DNA markers	https://shigen.nig.ac.jp/rice/oryzabase/	(Kurata and Yamazaki, 2006)
Panzea	Wheat SNP data	https://www.panzea.org/	(Zhao *et al*., 2006)
Pea Marker Database	V2 comprises 15,658 pea markers	www.peamarker.arriam.ru.	(Kulaeva *et al*., 2017)
PGDBj DNA Marker and Linkage Database	Large plant genomic marker base	http://pgdbj.jp/index.html?ln=en	(Asamizu *et al*., 2014)
PlantGDB	Maize RFLP markers	http://www.plantgdb.org/prj/	(Dong *et al*., 2004)
Phytozome	SNP and DNA marker database	http://www.phytozome.net/	(Goodstein *et al*., 2012)
Sol Genomics Network	Adam of Sodom, *Arabidopsis*, aubergine, pepper, potato, tobacco and tomato database with AFLP, DArT, INDEL, PCR, RFLP, RAPD, SNP and SSR markers	https://solgenomics.net/search/markers	(Fernandez-Pozo *et al*., 2015)
SSR Primer	An application that integrates SPUTNIK, an SSR repeat finder, with Primer3, a PCR primer design program to produce SSR sequences from a FASTA file	Retired	(Robinson *et al*., 2004; Jewell *et al*., 2006)
SSR Taxonomy Tree	A server with web-based searching and browsing of species and taxa for the visualization and download of these SSR amplification primers	Retired	(Jewell *et al*., 2006)

As genome assembly continued to become more common, genome assemblies from multiple individuals of the same species were compared and it became obvious that a single genome reference does not reflect the diversity of gene content between individuals of the same species. This led to the development of pan-genomic approaches for the identification of all genes for a species and the calling of gene presence/absence variation (PAV) between individuals (Golicz *et al*., 2015a). Initial pan-genomes used the assembly and comparison method, where two or more individuals were assembled *de novo* and compared. This is both expensive and suffers from the issue of false PAVs being called because of differences in assembly quality and annotation (Bayer *et al*., 2017, 2018), though it does have the advantage that the majority of genes are placed in a chromosomal location. Another approach is the read mapping and assembly method, which, when combined with SGSGeneLoss software, can produce a pan-genome assembly and call PAVs for very large numbers of individuals with relatively low sequence coverage. This approach has been successfully applied to several crop genomes, including *Brassica oleracea* (Golicz *et al*., 2016), bread wheat (Montenegro *et al*., 2017) and sesame (Yu *et al*., 2019). Because of the complementarity of the two methods, applying both would provide the most comprehensive pan-genome and understanding of gene PAVs across populations.

Application of Molecular Markers in Plant Breeding

Marker-assisted selection

Molecular markers can be used to improve the efficiency and quality of conventional plant breeding through the use of marker-assisted selection (MAS). MAS involves tracking DNA markers that are associated, and usually inherited, with a desirable trait, such as disease resistance, yield or abiotic stress (Collard and Mackill, 2008). MAS is advantageous because it reduces the time required to identify lines that express the specific traits, allowing field trials to focus on many complex traits that are poorly defined genetically and have no robust associated genetic markers. A prerequisite for MAS is an understanding of the inheritance of the trait, and genetic markers, which robustly associate with the trait in diverse backgrounds and environments. Traditionally, these markers have been developed using quantitative trait loci (QTL) studies; however, with the expansion of genetic, and particularly high-resolution genomic SNP data, linked markers are now mostly identified using genome-wide association studies (GWAS). GWAS, combined with large structure populations, such as nested association or MAGIC populations (Islam *et al*., 2016), have rapidly increased our understanding of which genes and allelic variants are contributing to agronomic traits in crops.

Genomic selection

Genomic selection (GS) is a marker-based selection technique that involves capturing the total genetic variance with genome-wide markers, based on genomic estimated breeding values (GEBVs) (Meuwissen *et al*., 2001). Unlike MAS and using QTL markers, GS uses all marker data as predictors of performance (Jannink *et al*., 2010), and the molecular and phenotypic data are used to calculate the GEBV for each marker (Crossa *et al*., 2017). GS tends to outperform MAS and has been shown to have great potential for plant breeding (Massman *et al*., 2013). In agriculture, GS can be used for accurately selecting parents for the genetic improvement of quality and yield traits without needing to phenotype every single individual, resulting in more rapid and lower cost selection (Jannink *et al*., 2010; Iwata *et al*., 2016). GS studies have so far been used on a variety of crops, including fruit trees, such as apple (Kumar *et al*., 2013), Japanese pear (Iwata *et al*., 2013) and grapevine (Fodor *et al*., 2014), as well as cereal crops, such as wheat (Massman *et al*., 2013), oats (Asoro *et al*., 2013) and maize (Beyene *et al*., 2015). GS has also been used for soybean breeding, where it was used to improve yield and agronomic traits in a breeding programme utilizing genotyping-by-sequencing. The programme reported high prediction accuracies, suggesting that GS could be used to improve grain yield (Jarquín *et al*., 2014).

Genome editing

With the expanding population and global climate change, there is an urgent requirement to

accelerate the breeding of advanced crops, with genomic and bioinformatics supporting these advances (Abberton *et al.*, 2015; Mousavi-Derazmahalleh *et al.*, 2019). Advances in genome editing have the potential to revolutionize crop breeding (Scheben *et al.*, 2017; Scheben and Edwards, 2017; Hu *et al.*, 2018; Scheben and Edwards, 2018a, 2018b), reducing the breeding cycle by several years and allowing the introduction of traits that are not present in current breeding material, without the introduction of foreign DNA and some of the associated societal issues surrounding this (Hartung and Schiemann, 2014; Scheben and Edwards, 2017).

Currently, the most popular gene-editing platform is the clustered regularly interspersed short palindromic repeats (CRISPR)/CRISPR-associated protein (Cas) system, which has been co-opted from the immune system of bacteria and archaea, and subsequently developed to provide targeted and precise gene editing for any region in any genome (Ishino *et al.*, 1987; Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013). The Cas protein, of which the most commonly used is the Cas9 nuclease, is targeted to a region of the genome by a single guide RNA (sgRNA), which is designed to be specific and unique to the target region (Gasiunas *et al.*, 2012; Jinek *et al.*, 2012). The Cas9 nuclease protein induces a double-stranded DNA break (DSB) at a specific site, and can induce either homology directed repair (HDR) to introduce specific DNA modifications, or non-homologous end joining (NHEJ) (Pacher *et al.*, 2007; Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013).

The flexibility of the CRISPR system has many applications in developing new crop varieties, as not only can it modify the regulation of genes of interest, but it can also provide targeted genetic modification, adding genes (hybrid or not) into the breeding pool or providing new combinations of genetic variation within crop genomes. It is expected that a novel crop breeding process, which integrates the collection of genomic, gene function, genetic variation and gene regulation data as essential first steps to crop breeding, will be established before the integration of CRISPR/Cas as a gene-editing system (Scheben and Edwards, 2017).

CRISPR has been used within several crop species, including tomato, tobacco, maize and rice, to disrupt genes or promoters (Jiang *et al.*, 2013; Ito *et al.*, 2015; Nekrasov *et al.*, 2017; Ueta *et al.*, 2017), knockdown genes (Soyk *et al.*, 2017) and to swap promoters (Shi *et al.*, 2017). These kinds of genetic modifications are capable of producing a wide range of desirable traits, including increasing disease resistance (Wang *et al.*, 2014; Ali *et al.*, 2015; Baltes *et al.*, 2015; Ji *et al.*, 2015; Chandrasekaran *et al.*, 2016; Nekrasov *et al.*, 2017; Peng *et al.*, 2017), increasing yield (Li *et al.*, 2013), increasing drought tolerance (Shi *et al.*, 2017), or the alteration of harvest timing (Soyk *et al.*, 2017) and fruit ripening patterns (Ito *et al.*, 2015). The CRISPR/Cas system can be used to create large-scale chromosome rearrangements and artificial recombination (Pacher *et al.*, 2007; Sadhu *et al.*, 2016; Ordon *et al.*, 2017); for example, in rice where it was capable of producing a stable and heritable ~170 kbp deletion that removed five biosynthetic genes involved in labdane-related diterpenoid synthesis (Zhou *et al.*, 2014). While most applications of CRISPR have occurred within domesticated lines, a recent study used CRISPR/Cas to confer monogenic agricultural traits onto stress-tolerant wild tomato accessions. The resulting progeny retained stress tolerance traits as well as larger fruit with higher nutritional content compared to wild types (Li *et al.*, 2018). CRISPR/Cas systems have large potential as mechanisms of crop improvement; however, the application of this technology relies on the accurate construction and annotation of genomes and pan-genomes as well as bioinformatics tools to identify candidate editing targets.

Future Directions

There are a growing number of databases and tools being developed to manage and provide access to the vast amount of information that is becoming available to breeders (Lai *et al.*, 2012b; Hu *et al.*, 2018). These include national initiatives in bioinformatics (Schneider *et al.*, 2017) as well as international species-specific systems, such as the wheat information system and the international rice informatics consortium (Scheben *et al.*, 2018). Integration and coordination of these databases continues with the requirement to access diverse data across multiple locations, and this type of coordination is essential to ensure

that breeders have access to the data required to accelerate crop breeding.

Real-time genotyping in the field

Crop breeding is increasingly reliant on genotype information. Current methods in plant genotyping require sample collection in the field and the transport of tissue samples to specialized facilities with relatively large-scale and expensive equipment. To provide breeders with information in the field, portable and streamlined genotyping platforms are required, as is the integration of genotyping results into decision support systems, which requires significant bioinformatic input. The most portable DNA sequencing platform currently on the market is the Oxford Nanopore MinION, which, as of May 2019, claims to provide up to 30 Gb of data per sequencing unit, from 10 min of prep time. The MinION platform produces relatively long read lengths of tens to hundreds of thousands of base pairs, and has been reported to have relatively high error rates of 12–35% (Ashton *et al*., 2014; Ip *et al*., 2015; Laver *et al*., 2015; Lu *et al*., 2016; Bowden *et al*., 2019), which necessitates the use of parallel sequencing using Illumina sequencing technology (Deschamps *et al*., 2016; Goodwin *et al*., 2016) to correct these reads to accurately call genomic variants. It is possible that innovation in this area may improve the MinION sufficiently or produce a parallel genotyping platform that is capable of large-scale genomic variant detection in a high-throughput manner.

To determine the genomic variation within field populations by genome sequencing, the sequencing reads are aligned to a reference genome, and SNPs and indels relative to the reference genome can be called using a number of tools, including SAMtools (Li *et al*., 2009) and GATK (DePristo *et al*., 2011). PAV is common within crop species, and this variation can also be called from the sequencing reads, often providing insights into disease resistance and stress tolerance (Golicz *et al*., 2016; Montenegro *et al*., 2017). The characterization of SNPs and PAV within field samples can then be integrated with gene function, pathogen resistance and historical breeding data to inform breeding and planting decisions in the field. Currently, the computational requirements of aligning reads to a reference genome, calling SNPs and PAV and then linking these to gene function data, is a significant and limiting resource to large-scale genotyping platforms, especially given the large size, relatively high homology and significant repetitive element of crop genomes (Morrell *et al*., 2011; Michael and VanBuren, 2015). In addition to crop genotyping, pest and pathogen genotyping could also be done in the field with technologies such as these, assisting in pathogen containment and providing another level of insight to crop management and breeding programmes (Dodds and Rathjen, 2010; Klosterman *et al*., 2016). To achieve accurate genotyping in the field, improvements in sequencing technology need to be combined with streamlined, fast and easy to use bioinformatics pipelines that can transform raw sequencing information into insights that farmers and breeders can directly apply to their decisions.

Deep learning and crop breeding

Deep learning is making rapid advances into the bioinformatics of crop improvement. Deep learning refers to a wide range of statistical methods to identify trends and patterns within large and complex datasets, such as satellite images, surveillance unmanned aerial vehicles (UAV) (e.g. drones) or high precision platforms, natural language processing or DNA sequence analysis, and sensing devices (Tardieu *et al*., 2017). Deep learning includes a substantial training process, during which the deep-learning system can extract features of interest to classify the data and achieve high accuracy in the prediction inference on previously unseen data (Rai *et al*., 2019). This training process is a significant undertaking, and to achieve high accuracy requires a relatively large amount of data that has been carefully examined for biases. Previous studies have focused mostly on leaf health phenotype assessment in crops, such as peach, apple and grapevine (Chéné *et al*., 2012; Sladojevic *et al*., 2016); melon infected with *Dickeya dadantii* (Pineda *et al*., 2018); lemon myrtle (Heim *et al*., 2018); wheat (Odilbekov *et al*., 2018); tomato (Mokhtar *et al*., 2015); 14 crop species infected with 26 different diseases (Mohanty *et al*., 2016); and to estimate disease severity for the apple black rot disease (Wang *et al*., 2017). The use of deep learning on these image datasets

aims to streamline and automate the process of plant phenotyping to maximize the amount of phenotyping data available for plant breeding purposes.

In addition to image analysis, deep learning has also been applied to genomic datasets, though to date, mostly focused in the human genome. Deep learning has been successfully applied to a SNP and small indel variant caller tool based on the statistical difference between the sequences clustered (Poplin *et al.*, 2018), to predict splice site frequency (Bretschneider *et al.*, 2018; Jaganathan *et al.*, 2019) and the characterization of functional effects of non-coding variants to predict the chromatin effects of sequence alterations (Zhou and Troyanskaya, 2015). In plants, the deep learning was applied to discover regulatory motifs, to predict RNA editing and alternative splicing, and to interpret genetic variants (Alipanahi *et al.*, 2015; Zeng *et al.*, 2016). It also was applied to predict RNA- and DNA-binding proteins profiles, enabling the depiction of more sophisticated regulatory mechanisms (Alipanahi *et al.*, 2015); and advances were made in the prediction of effectors and proteins in the apoplast, and proteins with mitochondrial targets in *Arabidopsis thaliana* and potato, for which the previously available tools had a poor performance (Zhang *et al.*, 2018). Deep learning is capable of expanding our understanding of genome structure and characterizing a previously unclassified or unannotated gene's function, all of which can assist in determining targets of plant breeding. It is through this deep understanding of genome function that targets for genome editing can be identified.

Other direct applications of deep learning are to assist the determination of genetic factors that underlie complex phenotypes of interest, and the prediction of phenotypes from genotypes. Deep-learning algorithms have already been applied to predict the phenotype from genotypes in wheat (Ma *et al.*, 2018), and to model the plant's growth patterns (Namin *et al.*, 2018), which can assist breeders to identify the most interesting lineages early in the plant's development. Also, when applied to field crops, the growth pattern analysis can predict the crop productivity based on its development (Namin *et al.*, 2018). Moreover, the phenotype-to-genotype analysis can be used to characterize growth patterns in multiple environmental conditions (e.g. soil, temperature, humidity and light), or how each genotype responds to diseases, indicating the genetic features that might be associated with a given phenotype. These future implementations have major applications within plant breeding, and will contribute to the functional classification of genes, enabling the acceleration of crop breeding programmes, identification of complex phenotypic traits, and the correlation between the phenotype and underlying genetic factors (Knecht *et al.*, 2016).

With the expansion of intelligent farming, vast amounts of data can be collected in real time in the field, through a variety of sensors installed across the plantation or by using the images collected by drones, airplanes and satellite (Tardieu *et al.*, 2017). Automated disease recognition in the field is one of the most applied features of deep learning tools; it has the potential to prevent a significant crop yield loss by the rapid identification of disease and pest infestation symptoms, allowing a fast response to prevent spreading. A few deep-learning tools for field analysis have been developed, such as a classification tool for radish infected with Fusarium wilt (Ha *et al.*, 2017), an automated identification tool of northern leaf blight-infected maize (DeChant *et al.*, 2017), a tomato disease and pest detector (Fuentes *et al.*, 2017), a moth detector in field images that can identify and classify the type into different subgroups (Ding and Taylor, 2016; Cheng *et al.*, 2017) and some more general pest detectors across several crop species (Wang *et al.*, 2017; Dawei *et al.*, 2019). In preparation for future applications, it was suggested a collaboration between UAV and unmanned ground vehicle (UGV), in which while the UAV collects the images for the deep-learning analysis, the UGV equipped with a robotic arm is capable of removing unhealthy plants and/or weeds (Bhandari *et al.*, 2017). These field recognition tools can be adapted to receive input from a large variety of devices, including mobile cameras, and become a feasible option for crop producers in developing countries and small farmers (Picon *et al.*, 2018). In addition, a few applications using deep-learning technologies have been designed to enable an accurate yield estimation, such as the prediction of rice grain yield from aerial images (Yang *et al.*, 2019), wheat biomass from UAV images (Lu *et al.*, 2019) and growth stage recognition of six crop plants with field image (Yalcin, 2018).

Overall, the precise crop diagnosis provided by the deep-learning tools enables the access to

informed decision making and can greatly reduce the usage of natural resources and misuse of pesticides and fertilizers, lowering the overall usage of these chemicals on the crop (Cheng *et al*., 2017). Accurate yield estimation and growth monitoring are valuable resources for the farmers to plan the harvest and management of the production, and also provide valuable information for breeders to design more productive crop lineages. Deep learning is a rapidly evolving technique with the ability to automatically extract features and analyse complex multi-dimensional datasets with high scalability to Big Data that can be applied to a wide variety of plant breeding aspects. Many breeding bottlenecks can be addressed by the use of deep-learning algorithms, and the growing research aims to make these algorithms even more precise to support the development of better, more adapted agricultural plants.

Summary

Crop breeding continues to accelerate, supported by advances in molecular and genomic data production and analysis. As the datasets continue to grow, more advanced bioinformatics tools and approaches will be required to support the accelerated production of improved crop varieties.

References

Abberton, M., Batley, J., Bentley, A., Bryant, J., Cai, H., *et al*. (2015) Global agricultural intensification during climate change: A role for genomics. *Plant Biotechnology Journal* 14, 1095–1098.

Ali, Z., Abulfaraj, A., Idris, A., Ali, S., Tashkandi, M., *et al*. (2015) CRISPR/Cas9-mediated viral interference in plants. *Genome Biology* 16, 238.

Alipanahi, B., Delong, A., Weirauch, M.T. and Frey, B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 33, 831.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., *et al*. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191.

Asamizu, E., Ichihara, H., Nakaya, A., Nakamura, Y., Hirakawa, H., *et al*. (2014) Plant Genome DataBase Japan (PGDBj): A portal website for the integration of plant genome-related databases. *Plant Cell Physiology* 55, e8.

Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., *et al*. (2014) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology* 33, 296.

Asoro, F.G., Newell, M.A., Beavis, W.D., Scott, M.P., Tinker, N.A., *et al*. (2013) Genomic, marker-assisted, and pedigree-BLUP selection methods for β-glucan concentration in elite oat. *Crop Science* 53, 1894–1906.

Baltes, N.J., Hummel, A.W., Konecna, E., Cegan, R., Bruns, A.N., *et al*. (2015) Conferring resistance to geminiviruses with the CRISPR–Cas prokaryotic immune system. *Nature Plants* 1, 15145.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 19, 421–422.

Batley, J. and Edwards, D. (2009) Genome sequence data: Management, storage, and visualization. *Biotechniques* 46, 333–336.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology* 132, 84–91.

Batley, J., Hopkins, C.J., Cogan, N.O.I., Hand, M., Jewell, E., *et al*. (2007a) Identification and characterization of simple sequence repeat markers from *Brassica napus* expressed sequences. *Molecular Ecology Notes* 7, 886–889.

Batley, J., Jewell, E. and Edwards, D. (2007b) Automated discovery of single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular genetic markers. In: Edwards, D. (ed.) *Plant Bioinformatics*. Humana Press, Totowa, New Jersey.

Bayer, P.E., Ruperao, P., Mason, A.S., Stiller, J., Chan, C.-K.K., *et al*. (2015) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theoretical and Applied Genetics* 128, 1039–1047.

Bayer, P.E., Hurgobin, B., Golicz, A.A., Chan, C.-K.K., Yuan, Y., *et al*. (2017) Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnology Journal* 15, 1602–1610.

Bayer, P.E., Edwards, D. and Batley, J. (2018) Bias in resistance gene prediction due to repeat masking. *Nature Plants* 4, 762–765.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2009) GenBank. *Nucleic Acids Research* 37, 26–31.

Berkman, P.J., Manoli, S., Mckenzie, M., Kubaláková, M., Šimková, H., *et al*. (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnology Journal* 9, 768–775.

Berkman, P.J., Skarshewski, A., Manoli, S., Lorenc, M.T., Stiller, J., *et al*. (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics* 124, 423–432.

Berkman, P.J., Visendi, P., Lee, H.C., Stiller, J., Manoli, S., *et al*. (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal* 11, 564–571.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., *et al*. (2015) Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Science* 55, 154–163.

Bhandari, S., Raheja, A., Green, R.L. and Do, D. (2017) Towards collaboration between unmanned aerial and ground vehicles for precision agriculture. *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping II, International Society for Optics and Photonics*, Anaheim, California, 8 May 2017, 1021806. DOI: 10.1117/12.2262049.

Bolser, D., Staines, D.M., Pritchard, E. and Kersey, P. (2016) Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods in Molecular Biology* 1374, 115–140.

Bowden, R., Davies, R.W., Heger, A., Pagnamenta, A.T., De Cesare, M., *et al*. (2019) Sequencing of human genomes with nanopore technology. *Nature Communications* 10, 1869.

Bretschneider, H., Gandhi, S., Deshwar, A.G., Zuberi, K. and Frey, B.J. (2018) Cossmo: Predicting competitive alternative splice site selection using deep learning. *Bioinformatics* 34, i429–i437.

Burgess, B., Mountford, H., Hopkins, C.J., Love, C., Ling, A.E., *et al*. (2006) Identification and characterization of simple sequence repeat (SSR) markers derived in silico from *Brassica oleracea* genome shotgun sequences. *Molecular Ecology Notes* 6, 1191–1194.

Carollo, V., Matthews, D.E., Lazo, G.R., Blake, T.K., Hummel, D.D., *et al*. (2005) GrainGenes 2.0. An improved resource for the small-grains community. *Plant Physiology* 139, 643–651.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A. P., Tang, H., *et al*. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345, 950–953.

Chandrasekaran, J., Brumin, M., Wolf, D., Leibman, D., Klap, C., *et al*. (2016) Development of broad virus resistance in non-transgenic cucumber using CRISPR/Cas9 technology. *Molecular Plant Pathology* 17, 1140–1153.

Chéné, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., *et al*. (2012) On the use of depth camera for 3D phenotyping of entire plants. *Computers and Electronics in Agriculture* 82, 122–127.

Cheng, X., Zhang, Y., Chen, Y., Wu, Y. and Yue, Y. (2017) Pest identification via deep residual learning in complex background. *Computers and Electronics in Agriculture* 141, 351–356.

Collard, B.C. and Mackill, D.J. (2008) Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transaction of the Royal Society of London B, Biological Sciences* 363, 557–572.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., *et al*. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823.

Consortium, *The Brassica rapa* Genome Sequencing Project (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43, 1035–1040.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., *et al*. (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science* 22, 961–975.

Dawei, W., Limiao, D., Jiangong, N., Jiyue, G., Hongfei, Z., *et al*. (2019) Recognition pest by image-based transfer learning. *Journal of the Science of Food and Agriculture* 99, 4524–4531.

DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E.L., Yosinski, J., *et al*. (2017) Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. *Phytopathology* 107, 1426–1432.

Depristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491.

Deschamps, S., Mudge, J., Cameron, C., Ramaraj, T., Anand, A., *et al*. (2016) Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. *Scientific Reports* 6, 28625.

Ding, W. and Taylor, G. (2016) Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture* 123, 17–28.

Doddamani, D., Khan, A.W., Katta, M.A.V.S.K., Agarwal, G., Thudi, M., *et al*. (2015) CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea. *Database* 2015, bav078.

Dodds, P.N. and Rathjen, J.P. (2010) Plant immunity: Towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics* 11, 539.

Dong, Q.F., Schlueter, S.D. and Brendel, V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research* 32, D354–D359.

Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., *et al*. (2009a) AutoSNPdb: An annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Research* 37, D951–D953.

Duran, C., Appleby, N., Vardy, M., Imelfort, M., Edwards, D., *et al*. (2009b) Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnology Journal* 7, 326–333.

Duran, C., Edwards, D. and Batley, J. (2009c) Molecular marker discovery and genetic map visualisation. In: Edwards, D., Hanson, D. and Stajich, J. (eds) *Bioinformatics*. Springer, New York, pp. 165–189.

Duran, C., Singhania, R., Raman, H., Batley, J. and Edwards, D. (2013) Predicting polymorphic EST-SSRs in silico. *Molecular Ecology Resources* 13, 538–545.

Edwards, D., Batley, J. and Snowdon, R.J. (2013) Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics* 126, 1–11.

Edwards, D., Zander, M., Dalton-Morgan, J. and Batley, J. (2014) New technologies for ultrahigh-throughput genotyping in plant taxonomy. In: Besse, P. (ed.) *Molecular Plant Taxonomy*. Humana Press, New York, pp. 151–175.

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., *et al*. (2015) The Sol Genomics Network (SGN) – from genotype to phenotype to breeding. *Nucleic Acids Research* 43, D1036–D1041.

Fodor, A., Segura, V., Denis, M., Neuenschwander, S., Fournier-Level, A., *et al*. (2014) Genome-wide prediction methods in highly diverse and heterozygous species: Proof-of-concept through simulation in grapevine. *PLoS ONE* 9, e110436.

Fuentes, A., Yoon, S., Kim, C.S. and Park, S.D. (2017) A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* 17, E2022.

Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciemce of the USA* 109, E2579–E2586.

Golicz, A.A., Batley, J. and Edwards, D. (2015a) Towards plant pangenomics. *Plant Biotechnology Journal* 14, 1099–1105.

Golicz, A.A., Bayer, P.E. and Edwards, D. (2015b) Skim-based genotyping by sequencing. In: Batley, J. (ed.) *Plant Genotyping*. Springer, New York, pp. 257–270.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., *et al*. (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* 7, 13390.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., *et al*. (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* 40, D1178–D1186.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17, 333–351.

Ha, J.G., Moon, H., Kwak, J.T., Hassan, S.I., Dang, M., *et al*. (2017) Deep convolutional neural network for classifying Fusarium wilt of radish from unmanned aerial vehicles. *Journal of Applied Remote Sensing* 11, 042621.

Hane, J.K., Ming, Y., Kamphuis, L.G., Nelson, M.N., Garg, G., *et al*. (2017) A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: Insights into plant–microbe interactions and legume evolution. *Plant Biotechnology Journal* 15, 318–330.

Hartung, F. and Schiemann, J. (2014) Precise plant breeding using new genome editing techniques: Opportunities, safety and regulation in the EU. *The Plant Journal* 78, 742–752.

Heim, R.H. J., Wright, I.J., Chang, H.C., Carnegie, A.J., Pegg, G.S., *et al*. (2018) Detecting myrtle rust (*Austropuccinia psidii*) on lemon myrtle trees using spectral signatures and machine learning. *Plant Pathology* 67, 1114–1121.

Hu, H., Scheben, A. and Edwards, D. (2018) Advances in integrating genomics and bioinformatics in the plant breeding pipeline. *Agriculture* 8, 75.

Imelfort, M., Duran, C., Batley, J. and Edwards, D. (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal* 7, 312–317.

Ip, C.L. C., Loose, M., Tyson, J.R., De Cesare, M., Brown, B.L., *et al*. (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* 4, 1075.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. and Nakata, A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology* 169, 5429–5433.

Islam, M.S., Thyssen, G.N., Jenkins, J.N., Zeng, L., Delhom, C.D., *et al*. (2016) A MAGIC population-based genome-wide association study reveals functional association of GhRBB1_A07 gene with superior fiber quality in cotton. *BMC Genomics* 17, 903.

Ito, Y., Nishizawa-Yokoi, A., Endo, M., Mikami, M. and Toki, S. (2015) CRISPR/Cas9-mediated mutagenesis of the RIN locus that regulates tomato fruit ripening. *Biochemical and Biophysical Research Communications* 467, 76–82.

Iwata, H., Hayashi, T., Terakami, S., Takada, N., Sawamura, Y., *et al*. (2013) Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. *Breeding Science* 63, 125–140.

Iwata, H., Minamikawa, M.F., Kajiya-Kanegae, H., Ishimori, M. and Hayashi, T. (2016) Genomics-assisted breeding in fruit trees. *Breeding Science* 66, 100–115.

IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345, 1251788.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., Mcrae, J.F., Darbandi, S.F., Knowles, D., *et al*. (2019) Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548.

Jaiswal, P., Ni, J.J., Yap, I., Ware, D., Spooner, W., *et al*. (2006) Gramene: A bird's eye view of cereal genomes. *Nucleic Acids Research* 34, D717–D723.

Jannink, J.L., Lorenz, A.J. and Iwata, H. (2010) Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics* 9, 166–177.

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., *et al*. (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15, 740.

Jewell, E., Robinson, A., Savage, D., Erwin, T., Love, C.G., *et al*. (2006) SSRPrimer and SSR taxonomy tree: Biome SSR discovery. *Nucleic Acids Research* 34, W656–W659.

Ji, X., Zhang, H., Zhang, Y., Wang, Y. and Gao, C. (2015) Establishing a CRISPR-Cas-like immune system conferring DNA virus resistance in plants. *Nature Plants* 1, 15144.

Jiang, W., Zhou, H., Bi, H., Fromm, M., Yang, B., *et al*. (2013) Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Research* 41, e188.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., *et al*. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821.

Kaur, P., Bayer, P.E., Milec, Z., Vrána, J., Yuan, Y., *et al*. (2017) An advanced reference genome of *Trifolium subterraneum* L. reveals genes related to agronomic performance. *Plant Biotechnology Journal* 15, 1034–1046.

Keniry, A., Hopkins, C.J., Jewell, E., Morrison, B., Spangenberg, G.C., *et al*. (2006) Identification and characterization of simple sequence repeat (SSR) markers from *Fragaria* x *ananassa* expressed sequences. *Molecular Ecology Notes* 6, 319–322.

Klosterman, S.J., Rollins, J.R., Sudarshana, M.R. and Vinatzer, B.A. (2016) Disease management in the genomics era – summaries of focus issue papers. *Phytopathology* 106, 1068–1070.

Knecht, A.C., Campbell, M.T., Caprez, A., Swanson, D.R. and Walia, H. (2016) Image Harvest: An open-source platform for high-throughput plant image processing and analysis. *Journal of Experimental Botany* 67, 3587–3599.

Kulaeva, O.A., Zhernakov, A.I., Afonin, A.M., Boikov, S.S., Sulima, A.S., *et al*. (2017) Pea Marker Database (PMD) – A new online database combining known pea (*Pisum sativum* L.) gene-based markers. *PLoS ONE* 12, e0186713.

Kumar, S., Garrick, D.J., Bink, M.C., Whitworth, C., Chagne, D., *et al*. (2013) Novel genomic approaches unravel genetic architecture of complex traits in apple. *BMC Genomics* 14, 393.

Kurata, N. and Yamazaki, Y. (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiology* 140, 12–17.

Lai, K., Duran, C., Berkman, P.J., Lorenc, M.T., Stiller, J., *et al*. (2012a) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal* 10, 743–749.

Lai, K., Lorenc, M.T. and Edwards, D. (2012b) Genomic databases for crop improvement. *Agronomy* 2, 62–73.

Lai, K., Lorenc, M.T. and Edwards, D. (2015) Molecular marker databases. In: Batley, J. (ed.) *Plant Genotyping*. Springer, New York, pp. 49–62.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., *et al*. (1987) MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174–181.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., *et al*. (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* 3, 1–8.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, S., Zhao, B., Yuan, D., Duan, M., Qian, Q., *et al*. (2013) Rice zinc finger protein DST enhances grain production through controlling Gn1a/OsCKX2 expression. *Proceedings of the National Academy of Science of the USA* 110, 3167–3172.

Li, T., Yang, X., Yu, Y., Si, X., Zhai, X., *et al*. (2018) Domestication of wild tomato is accelerated by genome editing. *Nature Biotechnology* 36, 116–1163.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., *et al*. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications* 5, 3930.

Lorenc, M.T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., *et al*. (2012) Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology* 1, 370–382.

Love, C.G., Robinson, A.J., Lim, G.A., Hopkins, C.J., Batley, J., *et al*. (2005) Brassica ASTRA: An integrated database for *Brassica* genomic research. *Nucleic Acids Research* 33, D656–D659.

Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics* 14, 265–279.

Lu, N., Zhou, J., Han, Z., Li, D., Cao, Q., *et al*. (2019) Improved estimation of aboveground biomass in wheat from RGB imagery and point cloud data acquired with a low-cost unmanned aerial vehicle system. *Plant Methods* 15, 17.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., *et al*. (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., *et al*. (2013) RNA-guided human genome engineering via Cas9. *Science* 339, 823–826.

Marshall, D.J., Hayward, A., Eales, D., Imelfort, M., Stiller, J., *et al*. (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods* 6, 19.

Massman, J.M., Jung, H.-J.G. and Bernardo, R. (2013) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Science* 53, 58–66.

Matthews, D.E., Carollo, V.L., Lazo, G.R. and Anderson, O.D. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Research* 31, 183–186.

Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Michael, T.P. and VanBuren, R. (2015) Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology* 24, 71–81.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., *et al*. (2002) Characterization of the flanking regions of *Zea mays* microsatellites reveals a large number of useful sequence polymorphisms. *Theoretical and Applied Genetics* 105, 532–543.

Mohanty, S.P., Hughes, D.P. and Salathé, M. (2016) Using deep learning for image-based plant disease detection. *Frontiers in Plant Science* 7, 1419.

Mokhtar, U., Ali, M.A. S., Hassanien, A.E. and Hefny, H. (2015) Identifying two of tomatoes leaf viruses using support vector machine. In: Mandal, J.K., Satapathy, S.C., Sanyal, M.K., Sarkar, P.P. and Mukhopadhyay, A. (eds) *Information Systems Design and Intelligent Applications*. Springer, New Delhi, India, pp. 771–782.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., *et al*. (2017) The pangenome of hexaploid bread wheat. *The Plant Journal* 90, 1007–1013.

Morrell, P.L., Buckler, E.S. and Ross-Ibarra, J. (2011) Crop genomics: Advances and applications. *Nature Reviews Genetics* 13, 85.

Mousavi-Derazmahalleh, M., Bayer, P.E., Hane, J.K., Valliyodan, B., Nguyen, H.T., *et al*. (2019) Adapting legume crops to climate change using genomic approaches. *Plant, Cell & Environment* 42, 6–19.

Namin, S.T., Esmaeilzadeh, M., Najafi, M., Brown, T.B. and Borevitz, J.O. (2018) Deep phenotyping: Deep learning for temporal phenotype/genotype classification. *Plant Methods* 14, 66.

Nekrasov, V., Wang, C., Win, J., Lanz, C., Weigel, D., *et al*. (2017) Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Scientific Reports* 7, 482.

O'Sullivan, H. (2007) GrainGenes – A genomic database for Triticeae and Avena. In: Edwards, D. (ed.) *Plant Bioinformatics*. Springer, New York, pp. 301–314.

Odilbekov, F., Armoniené, R., Henriksson, T. and Chawade, A.C. (2018) Proximal phenotyping and machine learning methods to identify septoria tritici blotch disease symptoms in wheat. *Frontiers in Plant Science* 9, 685.

Ordon, J., Gantner, J., Kemna, J., Schwalgun, L., Reschke, M., *et al*. (2017) Generation of chromosomal deletions in dicotyledonous plants employing a user-friendly genome editing toolkit. *The Plant Journal* 89, 155–168.

Pacher, M., Schmidt-Puchta, W. and Puchta, H. (2007) Two unlinked double-strand breaks can induce reciprocal exchanges in plant genomes via homologous recombination and nonhomologous end joining. *Genetics* 175, 21–29.

Peng, A., Chen, S., Lei, T., Xu, L., He, Y., *et al*. (2017) Engineering canker-resistant plants through CRISPR/Cas9-targeted editing of the susceptibility gene CsLOB1 promoter in citrus. *Plant Biotechnology Journal* 15, 1509–1519.

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., *et al*. (2018) Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Computers and Electronics in Agriculture* 161, 280–290.

Pineda, M., Pérez-Bueno, M.L. and Barón, M. (2018) Detection of bacterial infection in melon plants by classification methods based on imaging data. *Frontiers in Plant Science* 9, 164.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., *et al*. (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* 36, 983.

Rai, A., Yamazaki, M. and Saito, K. (2019) A new era in plant functional genomics. *Current Opinion in Systems Biology* 15, 58–67.

Robinson, A.J., Love, C.G., Batley, J., Barker, G. and Edwards, D. (2004) Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* 20, 1475.

Ruperao, P. and Edwards, D. (2015) Bioinformatics: Identification of markers from next-generation sequence data. In: Batley, J. (ed.) *Plant Genotyping*. Springer, New York, pp. 29–47.

Ruperao, P., Chan, C.-K.K., Azam, S., Karafiátová, M., Hayashi, S., *et al*. (2014) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant Biotechnology Journal* 12, 778–786.

Sadhu, M.J., Bloom, J.S., Day, L. and Kruglyak, L. (2016) CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science* 352, 1113–1116.

Savage, D., Batley, J., Erwin, T., Logan, E., Love, C.G., *et al*. (2005a) SNPServer: A real-time SNP discovery tool. *Nucleic Acids Research* 33, W493–W495.

Savage, D., Logan, E., Erwin, T., Robinson, A., Love, C.G., *et al*. (2005b) Real-time mining of maize genome data for simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) molecular genetic markers. *Maize Genetics Cooperation Newsletter* 50–51.

Scheben, A. and Edwards, D. (2017) Genome editors take on crops. *Science* 355, 1122–1123.

Scheben, A. and Edwards, D. (2018a) Bottlenecks for genome-edited crops on the road from lab to farm. *Genome Biology* 19, 178.

Scheben, A. and Edwards, D. (2018b) Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced allelic series to optimize quantitative and qualitative traits. *Current Opinion in Plant Biology* 45, 218–225.

Scheben, A., Wolter, F., Batley, J., Puchta, H. and Edwards, D. (2017) Towards CRISPR/Cas crops – bringing together genomics and genome editing. *New Phytologist* 216, 682–698.

Scheben, A., Chan, C.-K.K., Edwards, D., Mcnally, K.L., Mansueto, L., *et al*. (2018) Progress in single-access information systems for wheat and rice crop improvement. *Briefings in Bioinformatics* 20, 565–571.

Scheben, A., Verpaalen, B., Lawley, C.T., Chan, C.-K.K., Bayer, P.E., *et al*. (2019) CropSNPdb: A database of SNP array data for Brassica crops and hexaploid bread wheat. *The Plant Journal* 98, 142–152.

Schneider, M.V., Edwards, D., Gorse, D., Mcconville, M., Powell, D., *et al*. (2017) Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia. *Briefings in Bioinformatics* 20, 384–389.

Shi, J., Gao, H., Wang, H., Lafitte, H.R., Archibald, R.L., *et al*. (2017) ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnology Journal* 15, 207–216.

Shields, K., Ramsperger, M., Felitti, S. and Edwards, D. (2005) Endophyte ASTRA: A web-based resource for Neotyphodium and Epichloe EST analysis. In: Humphreys, M.O. (ed.) *Proceedings of the 4th*

International Symposium on the Molecular Breeding of Forage and Turf, a Satellite Workshop of the XXth International Grassland Congress. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 207–209.

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D. and Stefanovic, D. (2016) Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience* 2016, 3289801.

Soyk, S., Muller, N.A., Park, S.J., Schmalenbach, I., Jiang, K., *et al*. (2017) Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. *Nature Genetics* 49, 162–168.

Spangenberg, G., Sawbridge, T., Ong, E. and Edwards, D. (2005a) Ryegrass ASTRA: A web-based resource for *Lolium* EST analysis. In: Humphreys, M.O. (ed.) *Proceedings of the 4th International Symposium on the Molecular Breeding of Forage and Turf, a Satellite Workshop of the XXth International Grassland Congress*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 201–202.

Spangenberg, G., Sawbridge, T., Ong, E., Love, C., Erwin, T., *et al*. (2005b) Clover ASTRA: A web-based resource for *Trifolium* EST analysis. In: Humphreys, M.O. (ed.) *Proceedings of the 4th International Symposium on the Molecular Breeding of Forage and Turf, a Satellite Workshop of the XXth International Grassland Congress*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 195–196.

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. and Bennett, M. (2017) Plant phenomics, from sensors to knowledge. *Current Biology* 27, R770–R783.

Ueta, R., Abe, C., Watanabe, T., Sugano, S.S., Ishihara, R., *et al*. (2017) Rapid breeding of parthenocarpic tomato plants using CRISPR/Cas9. *Scientific Reports* 7, 507.

Van Ooijen, J.W. and Voorrips, R.E. (2001) JoinMap® version 3.0: Software for the calculation of genetic linkage maps. Plant Research International.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., *et al*. (2013) Draft genome sequence of kabuli chickpea (*Cicer arietinum*): Genetic structure and breeding constraints for crop improvement. *Nature Biotechnology* 31, 240–246.

Wang, G., Sun, Y. and Wang, J. (2017) automatic image-based plant disease severity estimation using deep learning. *Computational Intelligence and Neuroscience* 2017, 8.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., *et al*. (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology* 32, 947–51.

Ware, D., Jaiswal, P., Ni, J.J., Pan, X.K., Chang, K., *et al*. (2002a) Gramene: A resource for comparative grass genomics. *Nucleic Acids Research* 30, 103–105.

Ware, D.H., Jaiswal, P.J., Ni, J.J., Yap, I., Pan, X.K., *et al*. (2002b) Gramene, a tool for grass genomics. *Plant Physiology* 130, 1606–1613.

Weckx, S., Del-Favero, J., Rademakers, R., Claes, L., Cruts, M., *et al*. (2005) novoSNP, a novel computational tool for sequence variation discovery. *Genome Research* 15, 436–442.

Wilkinson, P.A., Winfield, M.O., Barker, G.L., Tyrrell, S., Bian, X., *et al*. (2016) CerealsDB 3.0: Expansion of resources and data integration. *BMC Bioinformatics* 17, 256.

Wilson, I.D., Barker, G.L., Beswick, R.W., Shepherd, S.K., Lu, C., *et al*. (2004) A transcriptomics resource for wheat functional genomics. *Plant Biotechnology Journal* 2, 495–506.

Yalcin, H. (2018) Phenology recognition using deep learning. *Electric Electronics, Computer Science, Biomedical Engineering Meeting (EBBT)*, Istanbul, Turkey, 18–19 April 2018. Institute of Electrical and Electronics Engineers, pp. 153–157.

Yang, Q., Shi, L., Han, J., Zha, Y. and Zhu, P. (2019) Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research* 235, 142–153.

Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., *et al*. (2011) Gramene database in 2010: Updates and extensions. *Nucleic Acids Research* 39, D1085–D1094.

Yu, J., Golicz, A.A., Lu, K., Dossa, K., Zhang, Y., *et al*. (2019) Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal* 17, 881–892.

Yuan, Y., Bayer, P.E., Lee, H.-T. and Edwards, D. (2017a) runBNG: A software package for BioNano genomic analysis on the command line. *Bioinformatics* 33, 3107–3109.

Yuan, Y., Bayer, P.E., Scheben, A., Chan, C.-K.K. and Edwards, D. (2017b) BioNanoAnalyst: A visualisation tool to assess genome assembly quality using BioNano data. *BMC Bioinformatics* 18, 323.

Yuan, Y., Milec, Z., Bayer, P.E., Vrána, J., Doležel, J., *et al*. (2018) Large-scale structural variation detection in subterranean clover subtypes using optical mapping. *Frontiers in Plant Science* 9, 971.

Zeng, H., Edwards, M.D., Liu, G. and Gifford, D.K. (2016) Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* 32, i121–i127.

Zhang, J.H., Wheeler, D.A., Yakub, I., Wei, S., Sood, R., *et al*. (2005) SNPdetector: A software tool for sensitive and accurate SNP detection. *Plos Computational Biology* 1, 395–404.

Zhang, N., Rao, R.S.P., Salvato, F., Havelund, J.F., Møller, I.M., *et al*. (2018) MU-LOC: A machine-learning method for predicting mitochondrially localized proteins in plants. *Frontiers in Plant Science* 9, 634. DOI: 10.3389/fpls.2018.00634.

Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., *et al*. (2006) Panzea: A database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research* 34, D752–D757.

Zhou, H., Liu, B., Weeks, D.P., Spalding, M.H. and Yang, B. (2014) Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Research* 42, 10903–10914.

Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* 12, 931.

7 Bioinformatics Approaches for Pathway Reconstruction in Orphan Crops – A New Paradigm

Dyfed Lloyd Evans[1,2,*] and Shailesh Vinay Joshi[1]
[1]*South African Sugarcane Research Institute, Durban, South Africa;* [2]*Cambridge Sequence Solutions (CSS), Waterbeach, Cambridge, UK*

Introduction

High-quality gene functional annotation is crucial for understanding the complex interplay between proteins in a plant's interactome. It is these interactions that ultimately yield a plant's phenotype, including their useful agronomic and domestication traits. Yet, despite the rapid advancements in genome assembly, gene annotation has generally lagged far behind. The situation is even worse when one considers orphan crops and species. Often, these species are evolutionarily distant from species whose genomes have been sequenced and few tools are directly available for their genome (or transcriptome) assembly and annotation. In standard crop nomenclature, an orphan crop (also known as a neglected or underutilized crop) is one that is not traded internationally but which may be of regional importance and, at best, receives limited attention by researchers. Several other crop species, although they may be economically important, can also fall under the umbrella of an orphan crop, at least in a genomic context. This is attributable either to a lack of genome sequencing, genome assembly, or to the availability of only a partial genome assembly that is associated with that particular species. Groups working on orphan crops and species are typically more resource-constrained than those working on 'major' crops. Even though these crops often play an important role in food security, income generation and supporting the nutritional needs of many developing countries, they are often ignored as being of regional importance only. However, in a broader context, they are often a good and underutilized source of resistance or tolerance to many biotic and abiotic stresses. However, many of these issues could be addressed if low-cost tools and techniques were available to analyse these crops at the genomic level. Though many different second-generation (short-read) and third-generation (long-read) sequencing platforms are available, they are costly, and analysing genomic data (assembly and annotation of transcriptomes and genomes) forms a major bottleneck in all species because of cost and the computational complexity of the assembly problem. High-quality gene functional annotation can help not only to decipher the functional genome of these crops, it will also attract funding for gene identification, with associated understanding of pathways and systems that are strongly associated with agronomic traits – delivering more focused breeding efforts. However, for most groups working on orphan crops, this annotation needs to be delivered with commodity hardware, or using publicly accessible servers.

* Email: dyfed.sa@gmail.com

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

In this chapter, we examine the gene assembly and annotation process, in particular as it applies to orphan crops and species and illustrate our methodologies utilizing, as an example, the construction of a grass lignin biosynthesis and polymerization pathway – the gateway to biofuel and bioprocessing.

The Genome Assembly and Annotation Process

Second-generation (short-read) sequencing technologies triggered an explosion of available genomic and transcriptomic resources in plant sciences (Bolger *et al.*, 2004). This phenomenon is set to continue and accelerate, as third-generation (long-read) sequencing platforms make it more efficient and cost effective to sequence and assemble even the most complex (highly polyploid and highly repetitive) plant genomes (Jiao *et al.*, 2017). This, however, has led to an annotation lag, as many plant genomes are not annotated manually but instead are functionally annotated on the basis of previously annotated genomes.

In the plant world, the one big exception to this phenomenon is the genome of thale cress, *Arabidopsis thaliana*. In Arabidopsis, The Arabidopsis Information Resource (TAIR) (Lamesch *et al.*, 2011) integrates community-based curations together with annotations gleaned from literature evidence. More than 2800 experimentally supported annotations have been included into the system between 2013 and 2015 alone (Berardini *et al.*, 2015). This wealth of curated data has been adopted and further augmented by AraPort (Krishnakumar *et al.*, 2014), an open-source resource, which encourages the community to contribute not only data modules but also visualization tools and apps. Despite these extensive resources, published data (Lamesch *et al.*, 2014) indicate that only about 77% of the protein-coding sequences could be assigned to any kind of structured annotation and this is the *best annotated* of all assembled plant genomes. Analysis of the latest annotation associated with the Arabidopsis genome within Ensembl plants (Bolser *et al.*, 2016) shows that this figure has hardly changed, with 79% of protein-coding genes now having functional annotation associated with them.

Compare this with the human genome annotation project, where, of the 20,418 total protein-coding genes as of EnsEMBL release 94, 19,033 genes (93%) have been manually annotated by the Sanger Centre HAVANA group and the National Center for Biotechnology Information (NCBI). These are available from the CCDS database release 22 (June 2018) (NCBI CCDS, 2018).

At the other end of the scale, we have the genome of *Sorghum bicolor* BTx623 (Paterson *et al.*, 2009), which, as of EnsEMBL plants release 41, has 34,118 protein-coding genes, of which only 1600 genes (4.7%) have associated functional annotation. This situation may change rapidly, however, as the NCBI have taken over annotation of this genome using their Gnomon pipeline (Sauvorov *et al.*, 2010).

The situation for orphan crops and species

For orphan crops and plants, whole transcriptome datasets, particularly using Illumina data, have become a surrogate for whole genome data and have been employed as a shortcut to a 'functional genome' (Hirsch *et al.*, 2013). However, assembly of these transcriptomic datasets is not a trivial task, despite these datasets being typically smaller and consisting uniquely of gene-rich data. Assembly accuracy is highly dependent on sequence depth, particularly for discovering rare genes. Moreover, the most popular transcriptome assembly tools, such as Trinity (Grabherr *et al.*, 2011), require significant optimization to produce an assembly of reasonable quality. Combining separate *de novo* assemblies from different assemblers may be one solution to improving such assemblies (Nakasugi *et al.*, 2014).

However, third-generation sequencing of transcriptomic datasets is overcoming some of the major limitations of short-read data (though error correction of long reads is still an issue). Third-generation sequencing data are dramatically changing our view of plant transcriptomes and genomes. For example, a survey of the *S. bicolor* transcriptome using Pacific Biosciences (PacBio) single-molecule real-time long-read isoform sequencing revealed more than 11,000 novel splice isoforms and 2100 novel genes (Abdel-Ghany *et al.*, 2016).

Though genome assembly, gene discovery and transcript mapping have advanced by leaps and bounds during the past decade, it still surprises most scientists that ascribing biological function to the genes in a process known as 'functional annotation' lags far behind the state-of-the-art in genome assembly. Surprisingly, performing the genome annotation task to a high degree of accuracy remains challenging, despite the extensive accrual of knowledge about gene function in model and crop species. The reasons for this are many, but one major aspect is the way that scientific information is still disseminated. Research findings remain largely published through academic journals. Many of these are only available to subscribers behind paywalls, a model that started in the 18th century. Even if the identification of gene function is published in an open-access journal, the associated annotation has, by some means, to be discovered computationally. Given the volume of research published each year, traditional publication strategies effectively create a barrier to annotation discovery (Tennant *et al.*, 2016).

For orphan crops and species, the situation is often far more complex, as without a closely related template, gene assembly is an almost insurmountable problem. For transcriptome assembly, using existing transcripts from other species to guide transcript assembly breaks down at 30% divergence (Ungaro *et al.*, 2017). Despite some advances, *de novo* assembly of plant transcriptomes remains a challenging problem, with typically only 60% of reference transcripts being recovered. Our recent analysis of *Miscanthus sinensis* transcriptome data (Table 7.1) shows that the situation is even more complex if an organism is polyploid or has resulted from recent genome duplications.

Comparison of several *de novo* and guided assembly methodologies is based on a reference assembly of the *M. sinensis* cv. Andante using ONT MinION sequencing (Lloyd Evans, unpublished) with associated gene calls. Transcript datasets were those for all named transcriptomic datasets in NCBI's sequence read archive as of the first week of November 2018. A total of 65,291 transcript coding loci and 101,377 protein-coding transcripts have been annotated. This is the reference against which the *de novo* and mapping transcripts were compared. Analyses were based on integrated Trinity and SOAPdenovo *de novo* assemblies, protein-guided assemblies and genome-guided assemblies. Transcript totals are corrected for contaminating sequences and fragmented transcripts based on the genome reference. Transcripts were also collapsed into gene bins cased on the reference assembly.

Miscanthus sinensis underwent a genome duplication/hybridization event about 2.3 million years ago. As a result, its gene complement is larger than those of many other Andropogoneae, with many genes that are close duplicates of each other. This provides a true challenge for transcript assemblers and makes this a good real-world test case to properly examine different assembly strategies. During the annotation of the assembled genome, genome-guided transcript calls were merged with protein datasets for the Andropogoneae. This revealed six very long genes that had not been annotated in their entirety. Contaminating sequences currently include

Table 7.1. Comparison of the results of *de novo* and guided transcript assembly methodologies.

Assembly type	Total transcripts	Contaminating sequences	Fragmented transcripts	Corrected transcript total	Total genes
Integrated Trinity and SOAPdenovo *de novo*	60,679	10,247	20,633	36,653	34,558
Protein-guided (diamond + SPAdes)	80,932	42	2	80,890	52,931
Integrated *de novo* and protein-guided transcriptomes	91,179	10,247	0	81,753	53,774
Genome-guided Trinity	87,011	6,441	700	79,907	54,002
Genome mapping (Tuxedo suite)	82,071	32	241	81,851	61,302
Genome-guided assembly	97,355	37	27	97,297	63,008
Reference genome	101,377	37	6		65,291

genes captured by the nuclear genome from the chloroplast and mitochondrion. These are marked as 'contamination' for historical reasons but are correct and functional parts of the nuclear genome. All other contaminating sequences represent bacterial, viral and fungal contaminating transcripts that had no cognates in the assembled genome.

Analysing the results in Table 7.1, compared with the *M. sinensis* genome reference (Phytozome), the two *de novo* assemblers, despite being the current best in breed (Honaas *et al.*, 2016), seem to perform worst even when the data from Trinity (Grabherr *et al.*, 2011) and SOAPdenovo-trans (Xie *et al.*, 2014) are combined. Though a good number of transcripts are obtained, more than 10,000 of these are attributable to contamination from bacterial, viral and fungal sequences. This is not unexpected, especially in plants; however, without a reference, it is hard to remove these sequences, as plants have captured exogenous genes. Moreover, 20,000 transcripts were fragmented and, in the genome, these collapsed to just more than 7000 real transcripts. This left 36,653 'real' transcripts that collapsed down to 34,558 true genes. Thus, in our hands, analysis of the complex genome of miscanthus against *de novo* assemblies only identified just more than 50% of the total gene complement.

Our novel protein-guided approach seems to perform better. Here, all unique proteins from the Andropogoneae (both NCBI mining and genome sequences) were used to bait transcript reads with Diamond (Buchfink, 2014), which were then assembled with SPAdes (Bankevich *et al.*, 2012). This allows the entire protein space of species closely related to miscanthus to be used for gene assembly. The pipeline takes both the best assembly for extension and any potential secondary copies. This enables us to assemble multiple transcripts from the same gene as well as secondary copies of the gene. As the assembly is to a protein reference, contamination from other species is almost eliminated and as we employ an iterative assembly process until the transcript cannot be assembled further, the number of fragmented transcripts drops dramatically. In all, we were able to assemble more than 80,000 transcripts that resolved to almost 53,000 genes. This is 81% of the total gene complement. Combining the protein-guided data with the *de novo* data revealed a further 843 genes, indicating that the protein space in the Andropogoneae may not be completely sampled.

Genome-guided Trinity identified 79,907 transcripts and 54,002 genes, a significant improvement from *de novo* assembly. Transcript mapping to the genome with transcriptome components of the Tuxedo suite (HISAT2, Cufflinks, StringTie and Ballgown) (Pertea *et al.*, 2016) identified 81,851 transcripts and 61,302 genes. However, none of these tools performed as well as our genome-guided assembly (based on our own modified implementation of JGI's PERTRAN) (Shu *et al.*, 2003), which revealed 97,297 transcripts and 63,008 genes. Interestingly, the fragmented transcripts were more commonly produced by transcripts with very high coverage than those with low coverage. This is not surprising, as too high coverage induces sequencing errors into the assembly graph that can break the graph. Subsampling (i.e. taking a 50% or 30% slice from the total transcript set) would solve this problem, indicating that assembling a subsampled dataset, along with the main dataset, would significantly reduce the number of fragmented transcripts in the overall assembly. Template-based methods do not have this issue.

It can be seen from Table 7.1 that while *de novo* assemblers perform a 'reasonable' job of assembling the transcriptome for a genome, they fail when there are multiple similar copies of cognate genes in the genome and they massively under-sample the alternate transcriptome. Indeed, our data mapping of ONT MinION whole transcriptome data to the miscanthus genome indicate that all short-read assemblers and mappers underestimate the alternate transcript complement of a genome by almost a factor of two (Lloyd Evans, personal communication).

However, for many orphan crops, there are no genome references. It is clear that guided assemblies outperform *de novo* assemblies, and even for orphan crops, it is still possible to use protein-guided assemblies based on distantly related species. The combination of protein-guided and *de novo* assembly significantly outperforms *de novo* assembly alone even when analysing complex genomes, and it is possible to assemble a much larger representation of the gene complement of an orphan crop without the need for a reference genome. Indeed, as the protein space

for related species becomes more populated, the transcript calls improve. The main problems and pitfalls of transcriptomic assembly for orphan species are discussed further by Ungaro *et al.* (2017).

Illumina sequencing technologies, while cheap, are not the optimal solution for transcript and gene discovery in orphan species. PacBio SMRT RNA-Seq technology offers a potential solution. Applications in orphan plants allow for the discovery of reference transcripts, alternate transcripts and UTR isoforms with a single analysis pipeline (Kim *et al.*, 2019). There is a caveat, however, as PacBio data are noisy, with a high error rate, and Illumina data are often required to polish and error-correct the PacBio reads prior to transcript binning. While applications for self-correction of PacBio RNA-seq data are being developed, the results, in our hands, have been unsatisfactory. Indeed, analysis of self-corrected PacBio transcriptomic datasets submitted to NCBI's TSA (transcript shotgun assembly) database (Coordinators, NCBI Resource *et al.*, 2014) reveals that very few of the transcripts have been assembled in their entirety, errors in transcript sequences still remain and few sequences translate cleanly. In addition, many datasets are heavily contaminated with bacterial, viral and fungal sequences.

Even with successful genome/transcriptome assembly and functional annotation, this is only the beginning of the story. Going back to the start of the efforts to assemble the human genome from 1998 onwards, at the time the below paradigm:

One gene → one transcript → one function

was still prevalent. In 2000, when Ewan Birney (now director of the European Bioinformatics Institute (EBI) in Cambridge, UK) started his human gene number sweepstakes, estimates of the human gene count ranged from 26,000 to more than 300,000 genes (with an average at about 40,000) (Willyard, 2018). At the time, the idea that 'increased complexity = increased gene count' was still prevalent. Today, Ensembl identifies and annotates 20,418 protein-coding genes and recognizes a total of 22,107 non-protein coding genes in the human genome. However, the number of protein-coding transcripts in the human genome has ballooned to 206,762 (http://www.ensembl.org/Homo_sapiens/Info/Annotation). Thus, the search for the origins of complexity has moved to the regulome (the regulatory network) and the interactome (protein–protein interactions within the cellular and extracellular compartments) of the organism.

Discoveries in the human genome typically inform genome studies in other organisms. This implies the need for increased efforts in discovering alternate transcripts and in deciphering the regulome and the interactome in plants.

Sugarcane

Sugarcane is a tall-growing true grass crop that is cultivated in tropical and sub-tropical regions of the world. It has the ability to store high concentrations of sucrose or sugar in its stem that, for this crop, is mostly consumed as refined sugar, along with other processed products, viz., ethanol, syrups, dextrans, confectionary, crude wax, glucose, jaggery (unrefined brown sugar cake), sugarcane juice and in the production of alcohol and spirits.

Sugarcane is a member of the genus *Saccharum* and belongs to the Andropogoneae tribe of the PACMAD clade of the Poaceae family of true grasses. Based on taxonomic and phylogenetic studies, sugarcane and its closely related interbreeding group of species have been ascribed to a group known as the *Saccharum* complex, which, itself, is a subset of the Saccharinae subtribe (Lloyd Evans *et al.*, 2019b). The *Saccharum* genus is characterized by its high ploidy level (6–12x) and sugarcane itself is an unbalanced hybrid, with a polyploid and aneuploid nature. Within the Andropogoneae, recent studies have revealed ancient network-style hybridizations (reticulation), indicating that many genera arose as the result of reticulate evolution (Welker *et al.*, 2015a; Lloyd Evans *et al.*, 2019b). These characteristics, along with poor taxon sampling and limited character matrices, have, traditionally, made it difficult to determine the true taxonomic relationships between Andropogoneae species, which has resulted in many conflicting taxonomic results (Daniels and Roach, 1987; Hodkinson *et al.*, 2002a; Sreenivasan *et al.*, 2015). However, more recent molecular analysis comparing genomic and chloroplastic data

has begun to more accurately resolve the systematics of *Saccharum* and allied genera (Lloyd Evans and Joshi, 2016; Lloyd Evans *et al*., 2019b).

Traditionally, the genus *Saccharum* was considered to be composed of six species: *S. spontaneum*, *S. officinarum*, *S. robustum*, *S. edule*, *S. barberi* and *S. sinense* (Daniels and Roach, 1987). Irvine (1999) proposed a revision to this categorization, with *S. robustum*, *S. edule*, *S. barberi*, *S. sinense* and *S. officinarum* all being folded into *S. officinarum*, leaving *S. spontaneum* as the only other species within the genus. His proposal was based on the inter-fertility of the grouped species and the lack of diagnostic characters to separate them into individual species at the molecular level (a taxonomic revision known as 'clumping'). Our whole-plastid genome study (Lloyd Evans and Joshi, 2016) revealed a new classification of *Saccharum* species. We proposed the division of *Saccharum* into four species: *S. spontaneum*, *S. robustum* (of which *S. edule* is a sterile cultigen), *S. officinarum* (wild type) and *S. cultum*. *Saccharum barberi* and *S. sinense*, being wild hybrids of *S. officinarum* and *S. spontaneum*, and thus not species in their own right, could be excluded on the basis of their hybrid nature alone. In addition, we inferred the lineage of modern hybrid cultivars to be derived from a cryptic founder species that was formally described and named as *S. cultum* (Lloyd Evans and Joshi, 2016). We demonstrated, phylogenetically, that *S. officinarum* and *S. cultum* lineages are descended from a female lineage that diverged from *S. robustum* about 750,000 years ago. The ancestors of modern and traditional sugarcanes, *S. cultum* and *S. officinarum*, diverged 650,000 years ago. We also confirmed that this divergence led to the separation of the classical old-world canes (*S. officinarum*) and the Polynesian canes (*S. cultum*), which gave rise to modern hybrid cultivars. We proposed a new model of sugarcane origins and a new pathway towards the molecular Saccharogenesis of sugarcane using true *S. officinarum* parentage.

The sugarcane genome is one of the most complex of all the crop genomes studied to date, and represents a major challenge for genomic studies because of its high degree of ploidy (12x) and highly repetitive sequence structure along with its aneuploid and interspecific (unbalanced hybrid) origin (D'Hont, 2005). Thus, sequencing a single copy of each cognate chromosome is insufficient to reveal the true genome complexity of the plant. Sequencing such a highly complex genome poses major challenges, especially because of its polyploid nature, the total genome size of sugarcane is approximately 1 Gb for a pseudo-monoploid genome and 12 Gb for the total genome size of modern hybrid cultivars, with chromosome numbers ranging from 100 to 130 (Grivet and Arruda, 2002; Souza *et al*., 2011).

The recent publication of a partial sugarcane draft genome sequence (Garsmeur *et al*., 2018) revealed that it contained only ~22,000 confirmed genes of the expected 72,000 genes in the sugarcane genome (sugarcane is an unbalanced hybrid). As sorghum was employed as an assembly and contig discovery template, few of the genes have validated functional annotation. Thus, in both a genomic and gene function context, sugarcane can be considered an orphan crop and can be employed as a model for genomic assembly and annotation strategies in other orphan crops.

Miscanthus

Miscanthus, in particular *Miscanthus* × *giganteus* is a prospective bioenergy crop with great promise. As a giant C4 grass, originally native to Asia, it displays high productivity at low input and is relatively cold tolerant (Jeżowski *et al*., 2017). This makes it suitable for growth in temperate Europe, China and North America; it is also suitable for marginal and reclaimed land, thus reducing its impact on food generation. Indeed, it has been estimated that in the USA, growing 11.8 million hectares (ha) of *M.* × *giganteus* would be required to produce 35 billion gallons of ethanol per year, while it would require 18.7 million ha of maize (corn) (both grain and stover) or 33.7 million ha of switchgrass to produce the same volume of ethanol (Anderson *et al*., 2011). This is marginally better than sugarcane in Brazil, where 12.4 million ha (both sugar and bagasse) would be required to generate the same volume of ethanol (Hofstrand, 2009).

As a triploid, *M.* × *giganteus* is sterile, meaning that it can be utilized in non-native areas without risk of escape. In addition, as it is grown from rhizomes, it is amenable to automated planting (Anderson *et al*., 2011). Miscanthus is 3.4 million years divergent from sugarcane and

is a member of the Saccharinae subtribe of the Andropogoneae (which also includes maize and sorghum) (Lloyd Evans *et al.*, 2019b).

Miscanthus × *giganteus* is believed to be a hybrid between diploid *M. sinensis* Anderss. and tetraploid *M. sacchariflorus* (Maxim.) Hack (Hodkinson *et al.*, 2002b). In its native Japan, *M. sinensis* is used for fodder and thatching, while in China, *M. sacchariflorus* is used for cellulose extraction. Each of the two parents is, itself, a hybrid of two ancestral *Miscanthus* species, with this hybridization event occurring about 1.8 million years ago (Lloyd Evans *et al.*, 2019b). Thus, the parental species have two copies of most lignin biosynthetic pathway genes and *M.* × *giganteus* has three. Understanding of lignification and secondary cell wall formation in the parent species will have dramatic implications for the breeding of *Miscanthus*, a grass just entering domestication, particularly as *M. sacchariflorus* is potentially cellulose rich.

Miscanes

Miscanthus lies within the natural hybridization window with *Saccharum* (Lloyd Evans and Joshi, 2016) and miscanes are hybrids between sugarcane and miscanthus (typically *M. sinensis* or *M. floridulus*). They were originally developed in South-east Asia in an attempt to transfer disease resistance genes from miscanthus to sugarcane. Disease resistance in these miscanes was initially apparent; however, an increase in biomass production was also noticed. For the hybrids, sugarcane contributes the characteristics of late flowering (allowing for more biomass production) with high yield, while miscanthus is responsible for cold tolerance, early spring sprouting, a more defined dormant period for harvesting and high vegetative yield. As a result, modern research has focused on the hybrid as a source of bioenergy (Somerville *et al.*, 2010).

Research indicates that miscanes share the growth potential and cold tolerance of miscanthus, allowing photosynthesis down to 10°C while retaining the ability to accumulate sugar in the stems, as in sugarcane. Miscanes are hardy to US zone 8 and can be grown on marginal land (Głowacka *et al.*, 2015). Understanding the lignin pathways in sugarcane and miscanthus will have implications for miscane development (Hoang *et al.*, 2015).

White fonio (*Digitaria exilis*)

Known in Senegalese French as fonio blanc, (white fonio – a term borrowed from *foño* in Wolof) and in The Gambia as *findi*, *Digitaria exilis* is a small grain 'millet' that may represent the oldest grain crop domesticated in West Africa, with cultivation starting around 7000 years ago (Garí, 2002). Fonio is one of the world's fastest growing cereals, reaching maturity in as little as 8 weeks and it is suitable for cultivation in areas with poor soils and low rainfall. However, the small grains are hard to dehusk, with the traditional method being to pound them in a mortar with sand. Typically, fonio is cooked as a pap (porridge) or it is directly steamed as a couscous substitute. It is also baked into flatbreads and used to make traditional beers (Jideani, 1999). However, the invention of a husking machine by Sanoussi Diakite, a professor of mechanical engineering from Senegal, has dramatically increased the efficiency of preparation, cutting the dehusking time from hours to minutes (Kumatoo, 2009), making this a more attractive crop across West Africa.

Digitaria exilis is classed as an 'underutilized crop' and is thus a true orphan crop. The plant is a member of the Paniceae tribe of the Panicoideae subfamily of grasses. The most closely related assembled genome is that of *Setaria italica* (foxtail millet) (Zhang *et al.*, 2012). Currently, within NCBI's sequence read archive (SRA), there are only four datasets in all, only one of which (SRR3938613) is large enough for assembly and mining. Compare this with more than 2000 datasets for sugarcane (*S. officinarum*/*Saccharum* hybrid), which is still considered an orphan crop, at least in a genomic context, and more than 6000 high-quality genome datasets for miscanthus.

Images of sugarcane, *M.* × *giganteus* and *D. exilis* are given in Fig. 7.1.

Lignin

Lignins are cross-linked phenolic compounds (Suslik, 1998) that are important structural materials in the support tissues of vascular

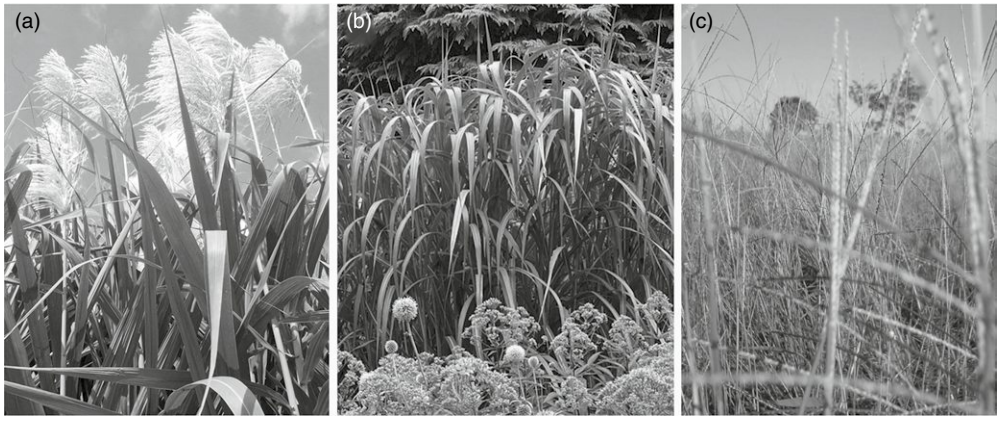


Fig. 7.1. Images of the three major crops discussed in this chapter. (a) Sugarcane photographed at the South African Sugarcane Research Institute. (b) *Miscanthus* × *giganteus* photographed at Knoll Gardens Nursery, Wimborne, UK. (c) *Digitaria exilis* photographed in Benin. (Images (a) and (b) copyright Dr D. Lloyd Evans, image (c) copyright M.M. Diop, with permission.)

plants. Lignins are crucial to the formation of plant's secondary cell walls, which provide additional protection to cells and rigidity and strength to the larger plant. The hydrophobic nature of lignin within these tissues is also essential to the containment of water within plant vascular tissues, allowing it to be conducted through the plant (Raven *et al.*, 2005). Vascular plants synthesize three main types of lignin monomers: sinapyl alcohol, S unit; coniferyl alcohol, G unit and *p*-coumaryl alcohol, H subunits (Liu *et al.*, 2018). However, though the core lignin biosynthesis pathway is common among vascular plants, there are differences in the pathways of monocots and dicots. Lignin comprises 30% of the typical lignocellulosic biomass of a plant, but the major current biorefinery systems almost all result in a lignin-containing waste stream. This makes the utilization of lignin for fungible fuels and bioproducts one of the most imminent challenges in modern biorefinery design (Xie *et al.*, 2016).

For second-generation biofuels and feedstocks for chemical processing, lignins are expected to play an increasingly important role. The sources of lignins will include agricultural by-products (straw or sugarcane bagasse), industrial process by-products (from paper making) as well as dedicated biofuel crops, such as switchgrass (*Panicum virgatum*), miscanthus (*Miscanthus* spp.), sweet sorghum (*Sorghum* spp.), eucalyptus (*Eucalyptus* spp.) and poplar (*Populus* spp.) (Welker *et al.*, 2015b). As such, lignocellulosic feedstocks are expected to play increasingly important roles in the bioeconomy.

For sucrose generation from lignin, increased S-lignin increases digestibility and sucrose yields (Li *et al.*, 2010). For secondary product generation, increased H-lignin content increases chemical and heat-based digestibility (Rinaldi *et al.*, 2016), though this is hard to achieve in grasses, which generally produce little H-lignin. For animal fodder, reduced overall lignin levels increase overall digestibility of both dry and fresh fodder (Chen *et al.*, 2004).

The lignin biosynthetic pathway is amenable to discovery by full text searching and pathway modelling. The enzymes can be assembled at the genomic, transcriptomic and structural levels. Moreover, altered expression of lignin biosynthesis and lignification genes can be examined at the gene/transcript expression level and the genes are amenable to transgenic manipulation. We demonstrate the reconstruction of lignification pathways in sugarcane (potential for second-generation ethanol production), *D. exilis* (potential for animal fodder and straw generation) and *M. sinensis* cv. Andante (potential for direct biomass production).

Conclusion

A new paradigm is needed for functional gene annotation and functional network discovery in

plants, particularly in orphan crops and species, where there are limited genomic/transcriptomic data available. We examine the current state-of-the-art in orphan species gene/transcript assembly and annotation as well as the construction of higher order interaction networks. These methodologies are illustrated using the lignin biosynthesis pathway in sugarcane, *M. sinensis* cv. Andante and white fonio (*D. exilis*) as examples.

Approaches to Assembling and Annotating Genes/Transcripts in Orphan Species

The orphan species gene/transcript assembly problem

Since the public release of NCBI's TSA database (https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA) in 2012, more and more *de novo* transcriptomic assemblies are being deposited in it every year. A curve representing the number of assemblies submitted to the TSA database from 2012 to 2018 is shown in Fig. 7.2.

Despite the dramatic increase in the number of transcriptomic datasets being made available, it could be argued that the quality of *de novo* transcriptomic assembly has not improved dramatically. Each transcriptome assembly algorithm has its own issues and the gold standard for transcriptomic assembly remains assembly against a reference (whether genomic or transcriptomic) (Honaas *et al.*, 2016).

Even long-read data have not yet significantly improved transcriptomic dataset assembly, as current long-read data contain significant sequencing errors and if not error-corrected, significant transcriptomic assembly errors occur (frame-shift mutations, missense mutations, coding errors, etc.). These make translation and annotation a challenge.

For template-based assembly strategies (mapping to a reference by read baiting and assembly based on a reference), assembly typically breaks down if there is >30% sequence divergence between the assembled and target sequences. This applies even to genome mappers that were specifically designed to map reads to a divergent reference, such as STAMPY (http://www.well.ox.ac.uk/project-stampy) (Lunter and Goodson, 2011). Assembly for all the methods above also breaks down if there is insufficient read depth and insufficient tissue sampling, meaning that rare transcripts are often not captured. Empirically, for good results, at least 4 Gb short-read sequences from whole seedlings, roots, stems, leaves and flowers should be sequenced and merged prior to any attempt at whole transcriptome assembly. This can be a real barrier to transcriptome assembly approaches in orphan crops and species, as funding for sequencing is often limited. However, emerging technologies, such as Oxford Nanopore's MinION technology (Lu *et al.*, 2016), which allows for whole-genome transcriptomics without the need for an amplification step at relatively low cost, seem likely to significantly reduce the transcriptome sequencing barrier for all organisms.

The above-mentioned issues pose a major problem for many orphan species, as, at a DNA level, they may be very divergent from the nearest sequenced/assembled relative. This divergence limit is also a general problem for genome annotation. To overcome this, annotators often work at the protein rather than the DNA level.

This approach has recently been applied to protein-coding gene assembly in Juncaceae and Pteridophytes (Lloyd Evans *et al.*, 2019a) genera, in which assembly against a reference was believed to be impractical, as they are too distant from their closest assembled reference genomes. The methodology relies on identifying groups of proteins orthologous to those to be assembled that are both evolutionarily antecedent and descendant to the species of interest. These proteins are placed in a FASTA-formatted file (multiple proteins can be placed in the same file). The protein file is indexed with Diamond (Buchfink, 2014) (available from: https://github.com/bbuchfink/diamond). Read files are converted to FASTA, and the FASTA sequences are then mapped to the protein files, again using Diamond. Scripts are used to sort the mapped read file identifiers into bins corresponding to the proteins mapped against. The application seqtk (https://github.com/lh3/seqtk) is then used to subset the FASTA read files based on the identifiers for each protein. These reads are then assembled. Even if the N- and C-terminal ends of the proteins have diverged significantly, a core transcript sequence is still obtained. This can then be used to extend the transcript sequence

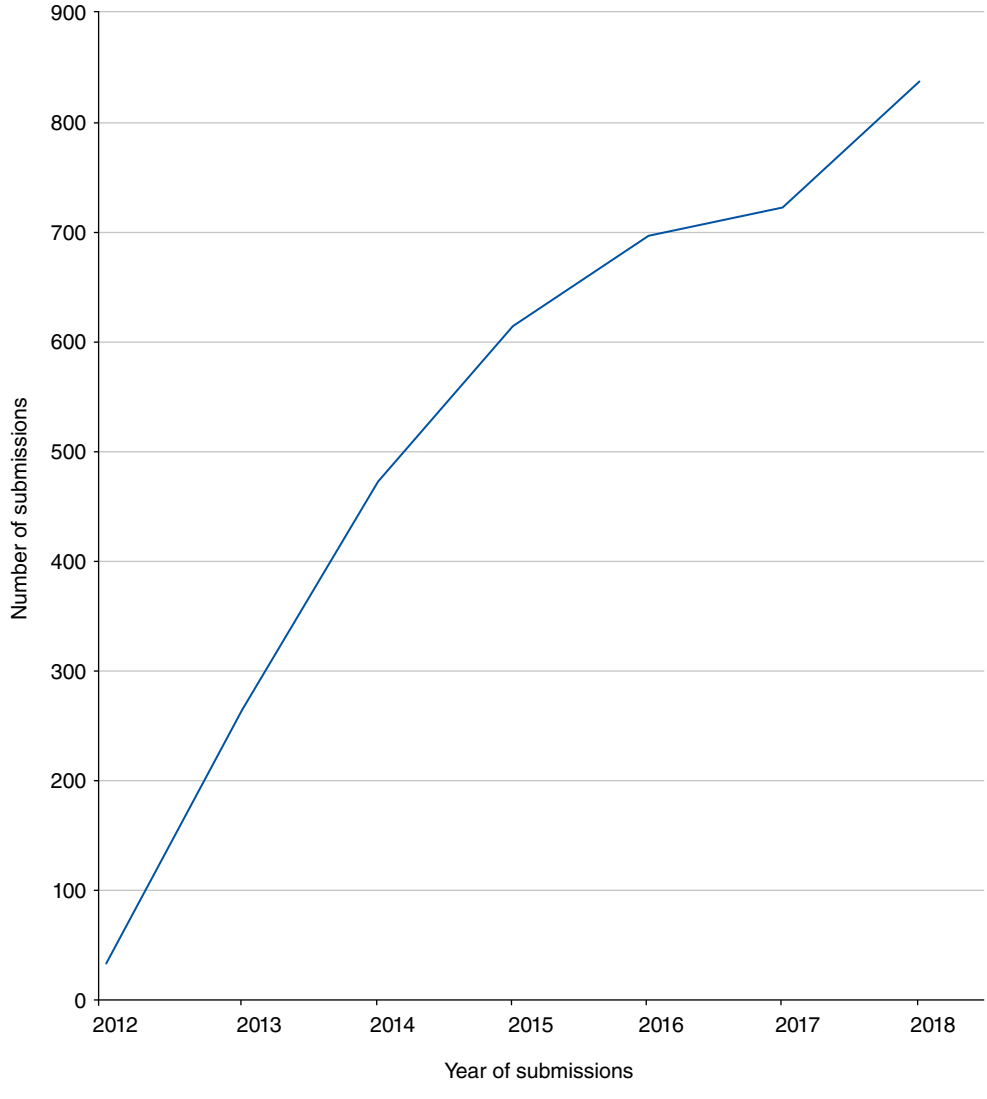


Fig. 7.2. Curve showing the trend of transcriptomic assembly submissions to NCBI's transcript shotgun assembly (TSA) database from 2012 to 2018. Despite a small apparent dip in 2017, probably due to all submitted sequences not having been released yet, there has been, essentially, a constant increase in the number of assemblies submitted to the NCBI throughout the database's lifespan.

using a standard bait-and-assemble methodology with multiple rounds until the sequence for the complete transcript is assembled.

Using this methodology, we have successfully assembled proteins from species that were >300 million years divergent – well outside the range at which DNA- or transcript-based reference guided assemblies are possible. Thus, the major barrier to reference-guided transcript assembly has been eliminated. The methodology is applicable to most orphan plant species. Moreover, the tool works on a laptop running Linux, even with large protein sets (hundreds of proteins) and short-read datasets of 25 to 30 gigabases (Gbp). This places reference-guided assembly within the reach of almost any group with a Linux-based laptop and a few terabytes of external storage.

The functional gene annotation problem

There are two main sources of functional gene annotation, i.e. primary and secondary. Primary annotation is based directly on functional studies and typically published in journal articles, whereas secondary annotation is based on protein/sequence homology, where the annotation from one gene in one species is transferred to a cognate gene in a second species. In most genome-annotation projects, this is done by BLAST comparison, or by employing gene functional ontologies. BLAST is most commonly performed against the nucleotide and protein datasets at the NCBI, though direct genome-to-genome BLAST analyses can also be used.

BLAST, however, has its own drawbacks. The algorithm is not symmetrical; thus, BLAST should always be performed query → target and target → query to ensure that the correct top hit has been identified. This is because BLAST is susceptible to query and target lengths and has a definite length bias. This also reveals another problem with NCBI BLAST, as the accuracy of matching is dependent on how populated the sequence space is within the NCBI database.

Consider the example given in Fig. 7.3, which shows how correct interpretation of BLAST analyses is critical to functional annotation. In this scenario, we have a six-domain protein of unknown function. This protein is orthologous to a partial two-domain sequence in NCBI. However, BLAST, as it is length-based, gives the top hit as an eight-domain protein because five of the domains in this protein are shared with our query – despite only sharing 90% identity. In fact, the eight-domain protein has a different function from our six-domain protein of interest. In many genome annotation pipelines, however, this top hit is taken naively and used to annotate the genome. This illustrates the pitfalls of annotation pipelines and explains why annotation remains a difficult problem without manual human intervention required to fine-tune annotation outputs. If, however, the two-domain protein had been BLASTed against both the six-domain protein and the eight-domain protein, then the correct match against the two-domain proteins would have been found. This is because BLAST is not a symmetrical algorithm and BLAST analysis in both directions (particularly when comparing short against long sequences) should always be performed to find the true best hit. As a result, many genome annotations found associated with real genome assemblies are of the type 'quite similar to an *A. thaliana* malic enzyme', as this is based on BLAST similarity. This is *correct* usage, as these types of annotations are liable to change upon reanalysis, but they often inflate the 'true' functional annotation percentage of a genome.

The thoroughness of this annotation process essentially depends on the existing knowledge about a gene function being transferable to genes of a similar sequence and assumes that this similarity reflects functional homology. Ensembl attempts to overcome this limitation by providing orthologue/paralogue data between different copies of a gene across genomes. When we know that an *Arabidopsis* gene is orthologous

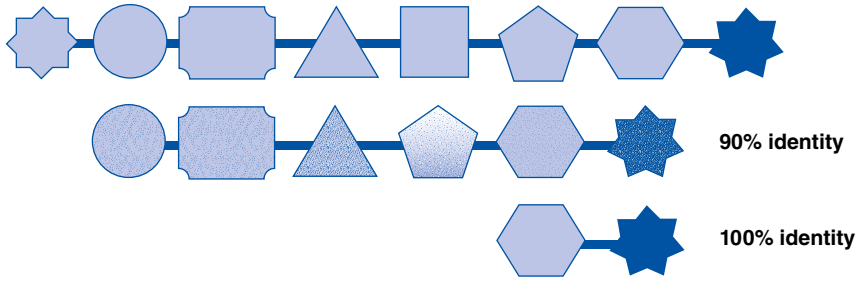


Fig. 7.3. Schematic domain-based diagram of a six-domain transcript blasted against NCBI's protein database. The order of the three proteins is that which a protein BLAST search would return. Each domain is given a different shape and the colour in the second protein is speckled to indicate that overall identity is 90% rather than 100%. In this example, the actual most similar sequence is a partial two-domain protein, but the blast top hit is against a more distantly related eight-domain protein as the overall match is longer.

to a cognate maize gene, for example, we have more confidence in ascribing the function of one gene to the other.

Even here there are pitfalls. In the example given in Fig. 7.4, we have a simulated evolutionary history of a single gene across six species, with an outgroup. There are two gene duplication events between species 2 and species 4. Species 5 has lost one of the new orthologous genes and species 6 has lost the copy of the original gene and one of the orthologous genes. Naively, gene 1 in species 6 looks like it is orthologous to gene 1 in all the other species. However, gene 1 is actually orthologous to gene 2 and is a paralogue of gene 1. It may be that gene 1 and gene 2 share the same function. But that is not necessarily the case and biochemical analyses would need to be performed to confirm this.

Thus, while evolutionary studies get us closer to the true orthologue of a given gene across species, the orthologue/paralogue issue needs to be carefully considered before ascribing a gene function by homology. One way to circumvent these problems is by the use of controlled vocabularies and functional ontologies (Hoehndorf *et al.*, 2015).

The Gene Ontology

Today, the Gene Ontology (GO) is by far the most widely used functional annotation ontology. It is implemented as a Directed Acyclic Graph (DAG), with a parent–child relationship between nodes in the graph. As a result, it is possible to infer more general terms from a specific term. As relationships form a hierarchical network, it is

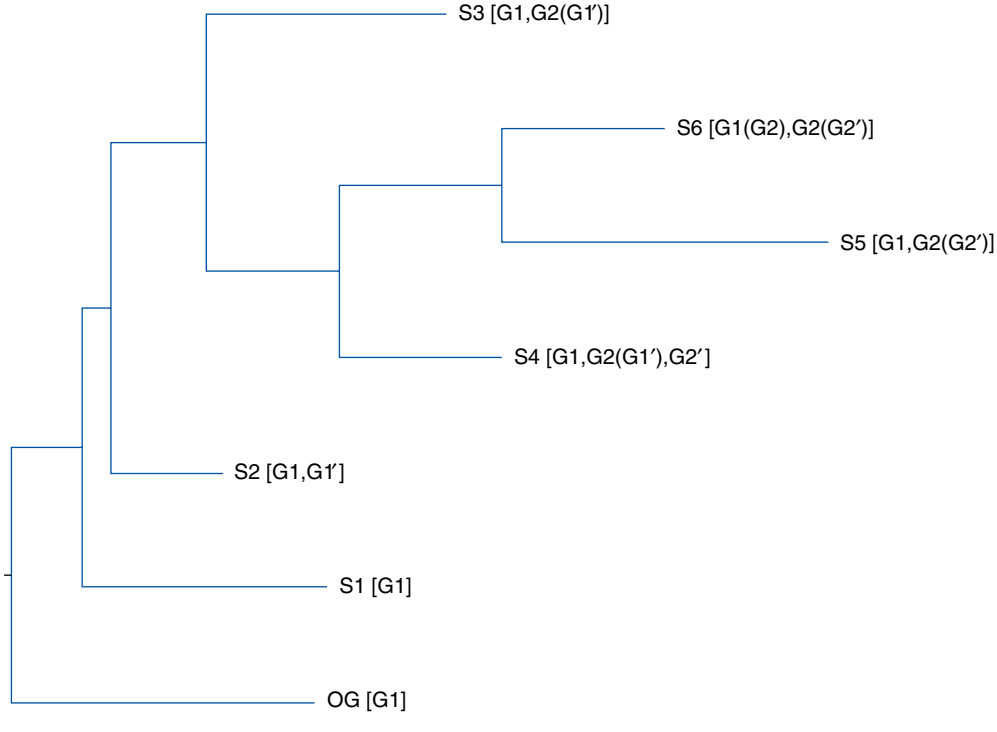


Fig. 7.4. Simulated evolution of six species (S1–S6) and an outgroup OG. The evolutionary history of a single gene and its duplicates (orthologues) is followed. The first gene name lists the primary or secondary gene copies, whilst the name in brackets follows the true name of the gene in terms of orthology. A prime (′) indicates a gene duplication. Species 5 has lost gene G2 compared with species 4. Species 6 has lost gene G1 and gene G2′ as compared with species 4. If the full evolutionary history of the gene were not known, genes G1 and G2 in species 6 might be taken to be the orthologues of genes G1 and G2 in species 4, whilst, in fact, they are the orthologues of genes G2 and G2′. This means that G1 in species 6 is an orthologue of gene G1 in species 4; they are not direct cognates of each other.

possible to group data in higher order parental annotation. For example, inputting a key lignin biosynthetic enzyme: 'hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase' (HC) into the GO yields the following functional class GO term: GO:0047172 (shikimate O-hydroxycinnamoyltransferase activity). Using the online AmiGO 2 browser (http://amigo.geneontology.org/amigo/) (Carbon *et al.*, 2008), reveals that the following gene: TAIR:locus:2154334 (AT5G48930) and the following additional terms: GO:0102660 (caffeoyl-CoA:-shikimate O-(hydroxycinnamoyl)transferase activity), GO:0047205 (quinate O-hydroxycinnamoyltransferase activity) and GO:0102093 (acrylate:acyl-coA CoA transferase activity) are associated with the gene query. Moreover, the child term is also associated with the following subcellular localization: GO:0005737 (cytoplasm).

The above illustrates the GO's three main annotation domains: 'Biological Process' – which describes the gene as a recognized series of events or molecular functions; 'Cellular Component' – describing the location of a biomolecule at a cellular and/or macromolecular level; and 'Molecular Function' – describing the function or abilities that a gene product possesses at the molecular level. In addition to GO terms, each GO annotation possesses an 'evidence code' as an attribute, which provides information on how a GO term was associated with a gene. Evidence codes indicate whether the annotation is based on experimental evidence, computational analysis, author statements or curatorial statements (each of these is manually curated). GO annotations also contain evidence codes, which are used to indicate whether the annotation is assigned by automatic/computational methods. This means that in a bioinformatics annotation pipeline, different types of evidence can be given different confidence scores.

The MapMan BIN Ontology

The MapMan BIN Ontology (Thimm *et al.*, 2004) is a plant-focused ontology that was originally developed to visualize 'omics, data on plant pathways. Subsequently, its function and scope have been extended and the current implementation (http://mapman.gabipd.org) contains about 2000 terms. In contrast to GO, MapMan is implemented as a hierarchical tree structure, with higher-level categories based on biological process and leaf categories containing detailed function. The database structure within MapMan was manually defined by experts in their respective fields; however, changes are applied periodically based on primary literature.

The Planteome

The Planteome (http://planteome.org/) (Cooper *et al.*, 2012) is a structured vocabulary and database resource that links plant anatomy, morphology and growth, and development to plant genomics data – offering a specific ontology for plants. Like the GO, it is accessed using AmiGO. Planteome is linked directly to GO, so it is possible to navigate through Planteome terms to GO terms. At the gene and functional levels, Planteome uses GO terms, but at higher levels, Planteome has plant-specific (PO) terms for plant anatomy and morphology. As a result, Planteome presents a plant-specific view sitting atop the GO ontology.

The plant interactome

The 'interactome' is defined as the complete set of molecular interactions in a particular cell. Primarily, this refers to direct interactions between molecules, such as protein–protein interactions, enzymatic interactions, protein–DNA interactions, protein–RNA interactions and DNA–DNA, RNA–RNA and DNA–RNA interactions. The term itself was first coined in 1999 by Bernard Jacq and colleagues (Sanchez *et al.*, 1999). It has also been hypothesized that the size of an organism's interactome is a better measure of the organism's complexity than its genome size or its total gene complement (Stumpf *et al.*, 2008).

Regarding experimental approaches, the two main methods to determine protein–protein interactions experimentally are the hybrid (Brückner, 2009) co-expression system and affinity capture mass spectrometry (Tureček, 2002). Both are high-throughput techniques and can be used as the basis for generating interaction networks.

Transcript expression profiling is also often used to study interactomes. In this case, up- and

downregulation of transcript in response to a stimulus (biotic and abiotic stresses, responses to diseases, parasites and predation, etc.) (Swindell, 2006), is measured experimentally.

Interactions, such as the ones described above, are typically viewed as networks defined relative to graph theory. Each interacting object is a node within the graph and the interaction is the vertex joining the nodes. Such interaction graphs can be highly complex, as shown in Fig. 7.5.

Generating interactome datasets requires extensive effort in proteomics, biochemistry and bioinformatics. However, they can be invaluable in assigning functions to novel genes. Indeed, such networks have been used successfully in yeast (Schwikowski *et al.*, 2000) and a range of 50 organisms (McDermott *et al.*, 2005) for gene functional annotation.

Transcript expression profiling has also been used extensively to determine stress response genes in plants, an example being chilling stress, response genes in *Oryza longistaminata* (Zhang *et al.*, 2017), or for whole-genome profiling of transcription factors in transgenic plants (Bond *et al.*, 2016).

Plant pathway databases

Pathway databases are repositories of metabolic, regulatory, transport, genetic, signalling and developmental pathways. In effect, a pathway is a subset of the plant interactome, but as they are handled separately in global databases, they are dealt with separately here.

Plant Reactome

Relative to breadth, Plant Reactome (http://plantreactome.gramene.org/) (Naithani *et al.*, 2016)

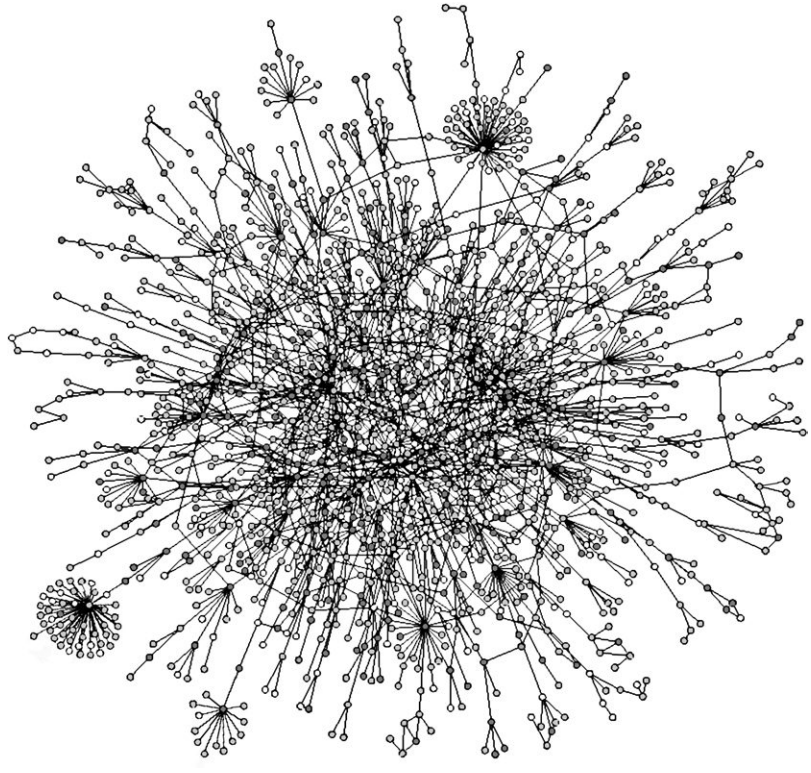


Fig. 7.5. Example plant interactome. Network diagram depicting two-protein interactions between 2000 *Arabidopsis thaliana* genes gleaned from full text searching of 2000 publications. Nodes are proteins and vertices (lines) are interactions. Some proteins lie at the centre of hubs and these are typically regulators, rate-limiting steps within biochemical pathways or transporter proteins.

is probably the most comprehensive database that covers metabolic, regulatory, transport, genetic, signalling and developmental pathways for 63 plant species, with *Oryza sativa* (rice) being the reference organism. Plant Reactome integrates data from 11 species/genus/group-specific databases: Gramene's Ensembl genome portal (Tello-Ruiz *et al.*, 2016), Phytozome (Goodstein *et al.*, 2011), PeanutBase (Dash *et al.*, 2016), TreeGenes (Wegrzyn *et al.*, 2008), Genome Database for Rosaceae (Jung *et al.*, 2013), SolGenomics Network (Fernandez-Pozo *et al.*, 2014), MaizeGDB (Andorf *et al.*, 2015), TAIR (Lamesch *et al.*, 2011), AraPort (Krishnakumar *et al.*, 2014), Legume Information System (Dash *et al.*, 2015) and Planteome (Cooper *et al.*, 2012), as well as data from six online resource providers: Gene Ontology (Hoehndorf *et al.*, 2015), EMBL-EBI's Gene Expression Atlas (Petryszak *et al.*, 2015), ChEBI (Hastings *et al.*, 2016), PubMed (Falagas *et al.*, 2008), UniProt (UniProt Consortium, 2014) and NCBI (ncbi.nlm.nih.gov). As of release 59, Plant Reactome covers 17,000 plant pathways, with a total of 42,000 interactions covering 89,000 genes. Given a protein name or identifier, Plant Reactome will map this to a function and a pathway, allowing enhanced functional annotation of a gene.

KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) is a knowledge base that allows for the systematic analysis of gene functions, linking genomic information with higher order functional information. The knowledge base consists of three main components:

1. Genomic information that is stored in the GENES database, a collection of gene catalogues derived from completely sequenced genomes and partial genomes that have up-to-date annotation of gene functions.
2. Higher-order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle.
3. A set of orthologue group tables that supplements the PATHWAY database, allowing information about conserved sub-pathways (pathway motifs) to be linked across organisms.

The knowledge base also has a LIGANDS database containing information about chemical compounds, enzyme molecules and enzymatic reactions. In general, pathways in KEGG are more broadly populated than other pathway databases and there are links through more species, so it is easier to find a true orthologue to your species of interest for functional annotation.

Gene ontology, along with KEGG (Kyoto Encyclopedia of Genes and Genomes), is used to conduct enrichment analysis for biological processes and pathways. KEGG is a database resource for understanding high-level functions and utilities of the biological systems. These includes the cell, the organism and the ecosystem, from molecular-level data, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. A useful functionality of KEGG is the KAAS (KEGG Automatic Annotation Server) (https://www.genome.jp/tools/kaas/), which provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database, including KEGG orthology term assignments and pathway localization. KASS works best when a complete set of genes in a genome is known. However, by querying amino acid sequences and the use of the SBH (single-directional best hit) method to assign orthologues, KASS can also be used for a limited number of genes (Moriya *et al.*, 2007).

PlantCyc

CYC is a suite of databases and software for pathway curation and visualization. The Pathway Tools are free for academics. The plant databases come under the PlantCyc umbrella (www.plantcyc.org) (Caspi *et al.*, 2007) and comprise a single multi-species reference database and 100 species/taxon-specific databases. PlantCyc is mainly used to describe enzymatic functions and allows for the construction of reaction networks. As the tools can be downloaded and deployed locally, it can be used to construct species-specific views on the metabolome and is particularly useful for reconstructing pathways in orphan species, but it must be coupled with some form of text searching to populate the databases (Nikoloski *et al.*, 2015).

Structural modelling

To get from a nuclear gene encoding for a protein to the functional form of that protein, several biological stages need to be traversed:

$$\text{DNA} \rightarrow \text{transcript} \rightarrow \text{protein}\,(1°) \rightarrow \text{folded protein}\,(2°) \rightarrow \text{higher order structures}\,(3°/4°)$$

Each of these stages contains different amounts and types of biological information. They also suffer different amounts of informational noise across evolutionary time. As, in the eukaryotic genetic code, 64 nucleotide triplets encode for 20 amino acids and stop codons, there is space for considerable sequence variation. As a result, highly divergent DNA sequences can code for very similar protein sequences. This is why, typically, proteins are preferred over DNA/RNA sequences for gene annotation.

Within proteins, only a subset of amino acids is responsible for the main structural domains of the protein (the secondary structure); a few more are responsible for the interactions that stabilize the secondary structural elements into the final tertiary structure of the protein. Indeed, it is possible for proteins with as little as 30% sequence identity to have the same three-dimensional structures (Tian and Skolnick, 2003) at least if the proteins are enzymes, enzyme inhibitors and structural proteins. This means that it is often easier to assign function and class to proteins (particularly enzymes) based on comparisons of three-dimensional structures rather than primary sequence comparisons.

Structural analyses and modelling used to be the domain of structural biologists with a physics background. Though interpretation of structural analyses is still complex, the barriers to entry have decreased dramatically during the past few years. Automated modelling portals, such as Imperial College, London's Phyre^2 service (http://www.sbg.bio.ic.ac.uk/phyre2/) (Kelley *et al.*, 2015), allow for a three-dimensional structure for a protein sequence to be predicted on the basis of existing structures within the Protein Data Bank (PDB) (Berman *et al.*, 2003).

In those cases where a high-quality modelling prediction is possible, Phyre^2 reports the confidence values, along with the structures used to construct the model and the function of those proteins. Thus, it is possible to use this methodology to ascertain protein function at the structural level.

Domain mapping

When sequence comparisons, GO mapping, pathway mapping and structural modelling fail to assign a function to a given protein, all is not lost. As protein secondary structure is intimately linked to protein function, it is possible to search protein secondary structural and protein domain databases to infer protein function. Table 7.2 contains a list of domain-focused databases and systems that can be used to help infer the function of a protein.

A tool that integrates searches across several of the above databases is MOTIF Search (https://www.genome.jp/tools/motif/). This tool also allows for the integration of a user-defined profile library (in PROSITE or HMMER) format into the search.

Even if domain assignment fails, all is not quite lost, as it is still possible to search for short, conserved motifs within a protein. One such application is the Eukaryotic Linear Motif (ELM) resource (http://elm.eu.org/) that scans input proteins for short linear motifs (SLiMs). These are then compared to a database of protein interaction sites composed of short stretches of adjacent amino acids. For the more advanced user, PDBeMotif (http://www.ebi.ac.uk/pdbe-site/pdbemotif/) can be used to explore the PDB to analyse structural motifs that can help place proteins in functional classes.

Understanding biological function

Before we proceed further, we need to define 'function' and 'functionality', particularly as they apply to biology. In English, the two words have multiple and overlapping meanings. However, in biology, there are two main concepts of function: the 'selected effect' and 'causal role'. 'Selected effect' (Neander, 1991) is an evolutionary term and implies that trait T has a proper biological function F that originated and is maintained reproductively. Therefore, a copy of a copy of a copy of T will still have the biological

Table 7.2. Resources for protein family and domain-based functional annotation.

Resource	Version	Number of families	URL	Comments
PFAM	32.0	17,929	https://pfam.xfam.org/	
TIGRFAMs	15.0	4,488	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	Not updated since 2014
PANTHER	13.1	15,524	http://pantherdb.org	
SMART	8.0	17,929	http://smart.embl-heidelberg.de/	Licence needed, free to academics
EggNOG	5.0	190,648 (37,127 plants)	http://eggnogdb.embl.de/#/app/home	
SCOPe	SCOPe 2.07-stable	4,919	http://scop.berkeley.edu/	Needs to be implemented locally
InterProScan	71	>40,000 integrated entries	http://www.ebi.ac.uk/interpro/	Meta search engine including all other resources (except EggNOG) but may not necessarily have the most recent version at all times
PROSITE	2018_10	1,229	https://prosite.expasy.org/	Allows a cartoon image of a protein's domains to be generated
CDD	3.16	56,066	http://www.ncbi.nlm.nih.gov/cdd/	Uses RPS-BLAST and includes partly older versions of PFAM, SMART and TIGRFAM

This table gives the domain identifying resource, the version as of this chapter's writing, the number of families indexed, the URL via which the resource can be accessed and some informative comments about the resource.

function *F*. The trait is under selection pressure and it is maintained because of its biological function *F*.

Contrast this with 'causal role', which is both ahistorical and non-evolutionary (Amundson and Lauder, 1994). Thus, for a trait, *Q*, to have a 'causal role' function, *G*, it is only necessary *and* sufficient that *Q* performs *G*. This is best and most simply illustrated by the case of two sequences in the genome. The first of these, TATAAA is a transcription-binding site. It has been conserved in the genome and maintained by natural selection to bind a transcription factor. Hence, it is a 'selected effect'. In contrast, consider a random genomic sequence TATAGA. This has no function per se. But in one generation, there is a mutation to TATAAA; now the sequence is the same as the transcription factor binding site. This sequence has what is called a 'selected effect function', which is to bind a transcription factor. However, this sequence has arisen merely by mutational happenstance and, purely by chance, it binds a transcription factor. But, as this is a random genomic sequence, transcription factor binding to the second sequence does not result in transcription; that is, it has no adaptive or maladaptive consequence. There can be no selection pressure to maintain the second sequence and it fits the 'causal role' model.

When annotating genomes, it is crucial to distinguish between annotatable elements (typically genes) that have a true function and those that only resemble functional elements. This is why determining evolutionary conservation is so critical in determining function.

The importance of phylogenetics

Phylogenetics, at both the species and gene levels, plays a crucial, but often overlooked, role in gene annotation. At the species level, it tells us how close two species or genera are to each other. The closer two genomes are in evolutionary terms, the more likely it is that one species can be used as an assembly template for the second species. It also makes it easier to infer the function of a protein in an unannotated species based on an

annotated protein in a reference species. Not only do we need to know the evolutionary history at the organismal level, we also need to know the evolutionary origin of proteins, particularly for large multi-protein families and this is the only way to properly classify the proteins. For example, consider the GST family of glutathione S-transferase enzymes, which are an important class of enzymes for herbicide detoxification in plants. There are two main categories of GSTs: phi and tau. However, it is often difficult to decide which class a particular protein lies within without performing phylogenetic analyses, as depicted in Fig. 7.6.

A close look at the phylogeny in Fig. 7.6 reveals that many of the proteins within the phylogeny have not been given an 'f' for phi or 'u' for tau designation; however, the phylogeny clearly separates phi and tau GSTs. Moreover, the phylogeny indicates that there are sub-groupings within the phi class GSTs that are yet to be explored and characterized. Thus, phi class GSTs should be divided into sub-classifications.

Also, as mentioned in the introduction, phylogenetic analyses reveal the relationship between cognate genes, revealing the orthologue (true homologue) and paralogue (related by gene duplication) inheritance of genes between species. Orthologues can gain gene annotation from another orthologue, but the function of paralogues may have changed and transfer of gene annotation between an orthologue and paralogue or between paralogues should be done with caution (unless there is additional supporting evidence for the functions to have remained the same).

A recent development in the evolutionary analysis of gene phylogenies for gene identification and annotation is the combination of phylogenetics and gene networks (Carvalho *et al.*, 2018), where the network approach aims to supplement and confirm/challenge the inferences made from a phylogenetic tree. This is particularly important in some enzyme classes, where a few amino acid changes can alter the function of an enzyme (e.g. the lignin enzymes PAL/TAL/PTAL; see section: The Sugarcane, Miscanthus and Fonio Blanc Lignin Pathways).

Defining orthology/paralogy is complex. One of the few general pipelines available is OrthoMCL (Li *et al.*, 2003), which can compare multiple species with one another. OrthoMCL can be invaluable for those working on orphan species, as complete genomes are not required for the algorithm to work. The latest implementation is available from orthomcl.org (http://orthomcl.org/orthomcl/) and the OrthoMCL wrapper code, which automates running the pipeline, is available from GitHub: https://github.com/apetkau/orthomcl-pipeline/. Once orthology to a reference gene in a reference genome has been established, the Ensembl plants database can be used to add more functional annotation. Computationally, OrthoMCL data can be captured in a MySQL database. In addition, the EnsEMBL orthology database is available to download and queries can be constructed across both systems using the EnsEMBL Compara system (Herrero *et al.*, 2016) and BioPerl.

Natural language processing

Rather than being associated with genomes, most gene functional annotation can be found in scientific literature published by groups working on the functional annotation of individual genes. As automated pipelines generally annotate genes based on public domain resources, the scientific literature represents a largely untapped resource of gene functional annotation. The same also applies to gene–gene interactions, gene co-expression studies, transcriptomic profiling studies and disease/mutant phenotypes.

Indeed, more than 400,000 journal articles were published in the biological sciences in 2017 alone (determined using an Entrez efetch query [https://www.ncbi.nlm.nih.gov/books/NBK25499/] at the NCBI). This is more than even a dedicated team of humans could parse in a reasonable amount of time.

One approach to reduce the complexity of the problem is to apply 'text mining' approaches. The simplest approach to this is full text search, where the entire corpus of a paper is placed in a database, and the text is searched using keywords via a system such as that implemented in MySQL (where the text is indexed prior to searching) (www.mysql.com) or Apache's SOLR/Leucene system (lucene.apache.org/solr/). While these solutions will allow text searching of documents, anyone who has queried Google Scholar (scholar.google.com) will know that for any full

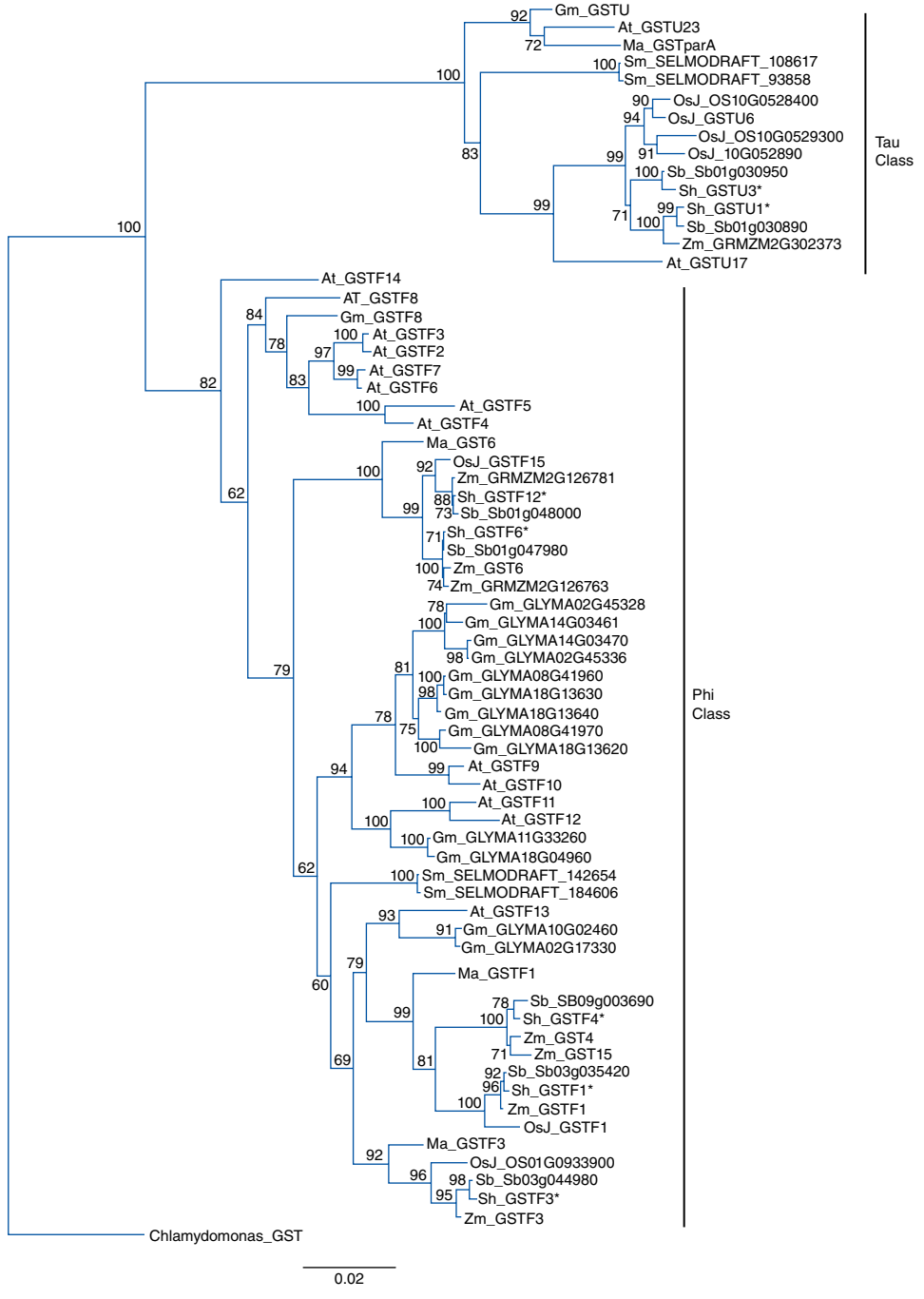


Fig. 7.6. Evolutionary history of the glutathione S-transferase (GST) multi-gene family. Phylogram for sugarcane tau and phi class GSTs with cognates from Sb – *Sorghum bicolor*; Sh – *Saccharum* hybrid; Zm – *Zea mays*; OsJ – *Oryza sativa* Japonica; Ma – *Musa acuminta*; At – *Arabidopsis thaliana*; Gm – *Glycine max*; Sm – *Selaginella moellendorffii* and Chlamydomonas – *Chlamydomonas reinhardtii*. The scale bar represents expected number of substitutions per site. Numbers above branches represent Maximum Likelihood branch support. The class to which a given protein belongs can only be accurately determined from phylogenetic analyses. (Adapted from Lloyd Evans and Joshi, 2017.)

text query, particularly using common terms, the false negatives may well outweigh the true positives.

The next step up from full text searching is 'natural language processing' or NLP, which is defined as the application of computational techniques to the analysis and synthesis of natural language and speech. NLP allows text corpora to be searched in a more naturalistic way, using language that scientists themselves would use in their publications. It is not just about searching for a limited number of keywords, but it is about the context and meaning of those keywords.

In scientific usage, NLP tasks can be divided into two main components: *Syntax* (breaking a corpus of text down into its syntactical components) and *Semantics* (this covers conversion of PDF documents to plain text, word sense disambiguation and question answering). For programmatic implementation, a full text searching system can be described, as in Fig. 7.7.

Implementing NLP for gene annotation

A system was developed based on Apache SOLR/leucene as the data storage engine and using the University of Sheffield's OpenNPL application, as implemented in the General Application for Text Engineering (GATE) developer framework (Cunningham *et al.*, 2013). For our particular use case (gene annotation and gene discovery), the GO and NCBI's taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) were also integrated into the overall system.

To reduce the size of the overall text corpus, initial searches were performed in Google Scholar (scholar.google.com). However, Google Scholar has no API (application programming interface) and queries had to be implemented in a scriptable web browser in a semi-automated way so that a user could solve one of the captcha images that Google Scholar regularly challenges a user with, particularly if many queries are made from a single computer IP.

All abstracts were downloaded from Google Scholar and where full text links were available, these were traversed to download the full text version of the article. PDF documents were converted to full text using the File::Extract::PDF Perl module. HTML documents were converted to full text with the HTML::Extract Perl modules (all modules are available from the Comprehensive Perl Archive Network [CPAN] www.cpan.org).

Article texts were imported into a SOLR database instance and were indexed. Subsequent to indexing, the following text processing steps were undertaken:

1. *Pre-processing*: tokenization, sentence boundary detection, lemmatization, part-of-speech tagging, species word identification, abbreviation detection and chunking.
2. *Named Entity Recognition*: this was performed with the GATE part of speech tagger.
3. *Term Normalization*: this involves choosing the correct identifier for each protein mention in the text, where the identifiers are drawn from a lexicon based on RefSeq. Fuzzy rules for matching were hand-written and species were normalized to the NCBI taxonomy.
4. *Relation Extraction*: to aid quick searching, positional relationships between words, particularly genes, interaction types, verbs, nouns and species names, were extracted, stored in a table and indexed. This included the current gene's part of speech tag as well as a lookup for the four preceding and following words. This was subsequently extended to the gene product, where known.

For querying, a novel NLP querying language was developed. In effect, this is a development shortcut, allowing for querying of a database in a natural language compatible way, without the need to write a parser for a natural language query.

An example is given below:

$$\begin{pmatrix} (\text{gene} \,|\, \text{transcript})[\text{n}] \\ \sim 5\left(\begin{pmatrix} \text{up-regulated} \,| \\ \text{down-regulated} \end{pmatrix}\right)[\text{adj}] \sim 5(\text{fold}) \end{pmatrix} \\ \pm 5(\text{heat} \,||\, \text{“heat stress”} \,||\, \text{drought}) \pm 5(\text{significant})$$

Converting this into English, the query is searching for the terms 'gene' or 'transcript' as a noun, which must lie within five words of the terms 'up-regulated' or 'down-regulated' used as an adjective. The terms used previously must all lie within five words upstream of the term 'fold' used as an adjective. The previous terms must have the word 'heat', the phrase 'heat stress' or the word 'drought' within five words upstream or downstream from them. All these

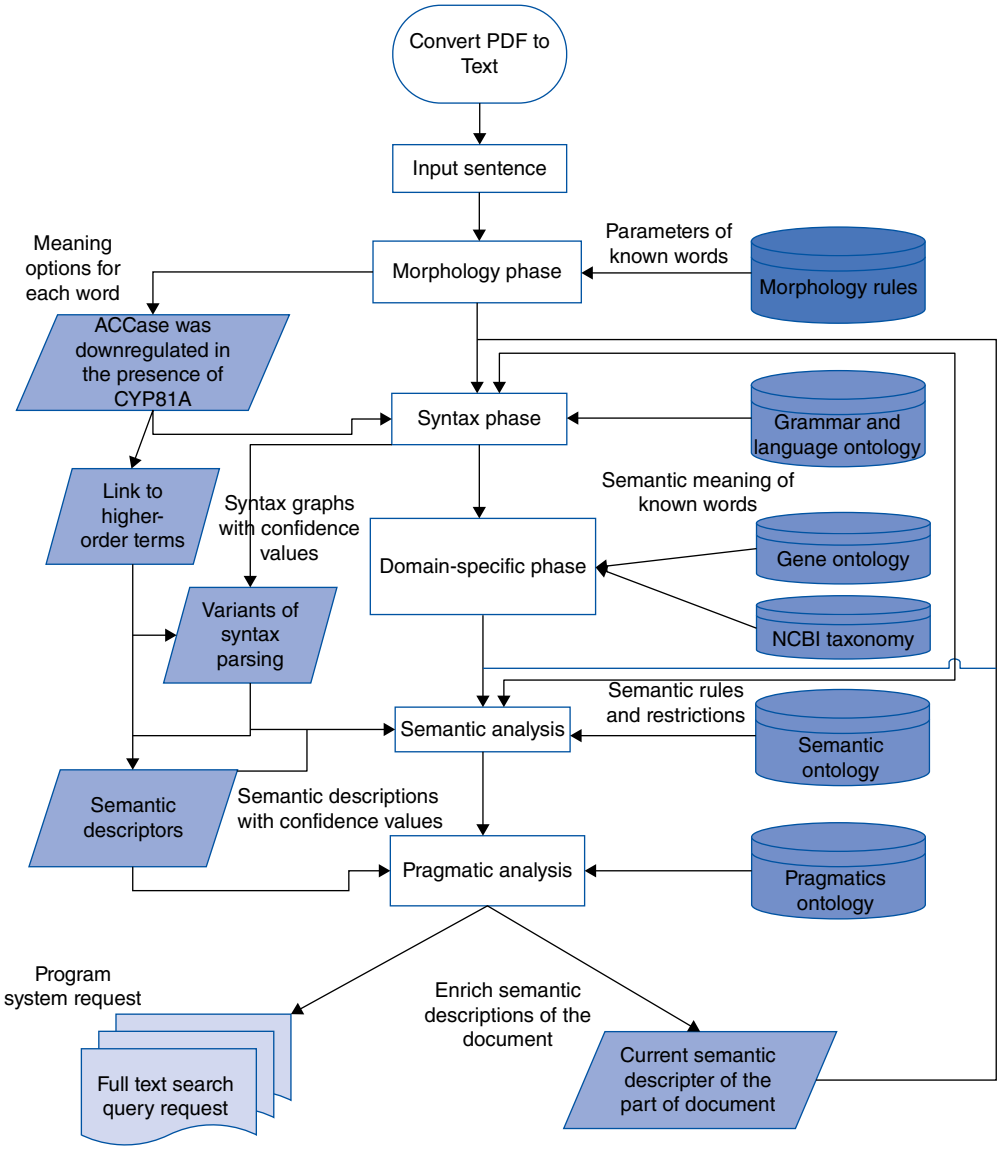


Fig. 7.7. Schematic representation of a Natural Language Processing Algorithm. The image represents the main steps required to develop a Natural Language Processing Algorithm. This algorithm is adapted to gene biology with the inclusion of the Gene Ontology (GO) and the NCBI taxonomy. The starting corpus being abstracts of published articles or article full text, typically in PDF format. Within the presented schema, ontologies are shown shaded on the right, operations are shown shaded on the left and the main central analysis steps are in white (in the centre).

terms are enclosed in round brackets, meaning that they must apply to a single sentence. Finally, the word 'significant' must lie within five words upstream or downstream of the previously described query, but it can lie in another sentence.

Opening and closing brackets group terms together as well as define which terms should co-occur within the same sentence (adding +2 +3 etc. to the end of round brackets means that the enclosed terms must line within two or three

sentences). If the query is enclosed in square brackets '[]' that means the terms apply to a paragraph, otherwise square brackets define parts of speech. This way, it is possible to construct very complex natural language type queries but only using logical constructs that are easy to parse.

This is a fairly lightweight implementation that works well for species-specific gene discovery and gene annotation. Apart from requiring terabytes of storage space, it can be implemented on a laptop rather than a server, thus reducing one of the main barriers to implementation.

For Java developers, an implementation, such as the one above, is probably the easiest to start with. However, for Python developers (which includes most bioinformaticians with a biology background), Python has a Natural Language ToolKit (NLTK) implementation, which can be installed using the Python package installer, Pip.

Beyond personal implementations, such as described above, the four state-of-the-art NLP tools are currently considered to be: MetaMap (Aronson, 2001), NCBO Annotator (Jonquet *et al.*, 2009), Textpresso (Müller, 2004) and SciGraph (Mungall *et al.*, 2016). However, it should be noted that these tools were developed to recognize and annotate pieces of text with ontology concepts. Thus, they are not gene functional annotation tools per se, but they can be used to extract articles from text corpora that are likely to contain gene annotation linked to ontology terms.

The big problem with natural language processing is that the language used by each domain of expertise is different. Therefore, while there are several commercial solutions for mining Internet data, form data and even biomedical text corpora, this is the extent of mature NLP implementations. This effectively leaves biological data out in the cold. As every domain of expertise needs its own lexicon – this is the basis for text mining – the only way to guarantee a domain-specific lexicon is to roll your own.

A short review of available commercial and free full text mining solutions is given below. These, therefore, focus on platforms that could be employed to construct a natural language processing engine that can be made aware of biological terms and meanings.

The Stanford Natural Language Processing group (https://nlp.stanford.edu/) has produced a suite of applications written in Java that encompasses text parsing, part of speech tagging, along with a classifier and a term-based extractor. If you are a Java developer, these components can quite readily be put together into a natural language processing and querying system.

1. Apache openNLP (https://opennlp.apache.org/) is fully open source and forms a toolkit for machine learning-based NLP that integrates well with Lucene and Solr.
2. Written in Python, spaCy (https://spacy.io/) is a fully featured NLP toolkit that allows developers to code industrial strength NLP applications.
3. NVIVO (https://www.qsrinternational.com/nvivo/home) is the first commercial product to make it to the list. Despite its steep learning curve, it performs free text parsing and association out of the box. It is not specifically designed for the bio domain but could be used as a first phase text-processing engine.
4. TextBlob (https://textblob.readthedocs.io/en/dev/) is a Python library for processing text data. For the bioinformatician, this probably provides the simplest introduction to NLP text processing, at least for common tasks in text processing.
5. Deeplearning4J (https://deeplearning4j.org/) is an open-source NLP implementation written in JVM/C++/Python that is intended for enterprise-scale applications. Once installed by the systems administrator, it also has interfaces that aim to enable fast prototyping for non-experts.
6. IBM's Watson Natural Language Classifier is a commercial product that aims to provide a service enabling developers without a background in machine learning or statistical algorithms to create natural language interfaces for their applications. It is very useful as a term classifier.

For a survey of other techniques, see Zeng *et al.* (2015). For a general introduction to Bioinformatics implementations of NLP, read the *Natural Language Processing with Python* book (Bird *et al.*, 2009).

The 'annotatable gene space'

Thus far, we have examined the various methods that can be employed to annotate genes and genomes. But the question remains – how much

of a plant's gene complement is actually annotatable and how much effort should be placed in gene/transcript annotation? The 'annotatable gene space' is a term often used by gene/genome annotators and it refers to the percentage of an individual genome's gene complement that can be annotated using public resources. Figure 7.8 represents the results of different annotation strategies relative to the effectiveness of whole genome gene annotation.

It is obvious from Fig. 7.8 that though GO, the Gene Ontology remains the go-to annotation resource for many genome annotators, it actually performs the worst, with 36% unannotatable genes. The situation is even worse if electronically derived data are considered unreliable. This means that GO fails to reliably annotate 69% of all Arabidopsis genes. Ensembl orthology mapping, where gene function is mapped based on orthologue relationships with the most closely related species getting precedence, performs the next best with 9% unannotatable genes and 30% unreliably annotated genes. At face value, domain analysis performs well, with only 8% unannotated genes, but an additional 55% of genes are annotated with low reliability. However, overall annotation quality is low and the annotation is at the domain level rather than the protein level. In this case, electronic (automated) annotation far outweighs experimental annotation.

In terms of the least percentage of unannotated genes and the least percentage (9%) of automated electronic annotation text mining by natural language processing outperforms the other methodologies by a large margin. However, implementing a text mining approach is complex and costly and requires an informaticians with extensive programming skills, who not only knows genomics, transcriptomics and

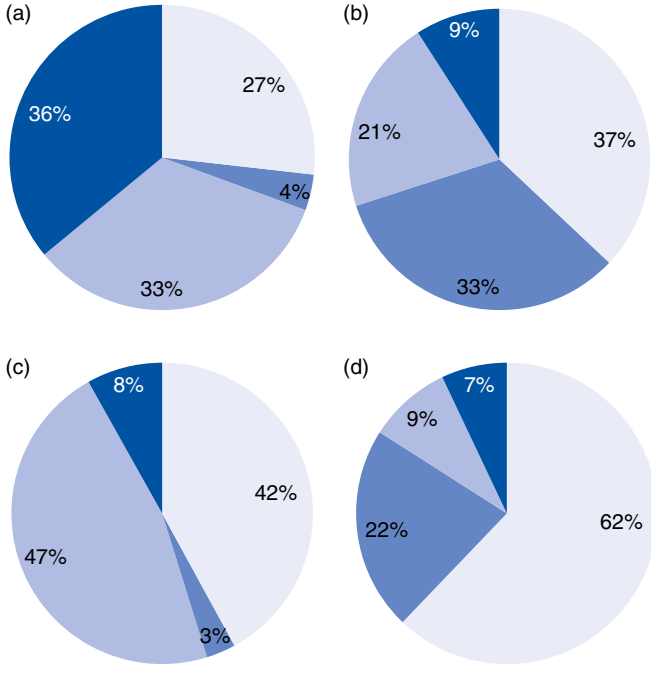


Fig. 7.8. The annotatable gene space. Evaluations of different gene annotation strategies. (a) Gene Ontology (GO) Molecular function and Biological Process annotation; (b) cross-species annotation using EnsEMBL orthology; (c) annotation via domain analysis; and (d) annotation via full text searching. (a), (b) and (c) were applied to the *Arabidopsis thaliana* genome data and (d) was applied to *Zea mays*. Key: the four shades from white to dark blue represent experimental data, human curated data, electronically derived data and unannotated genes and correspond to each respectively darkened shade. (Copyright Dr D. Lloyd Evans.)

genome annotation but who is also conversant with the principles of human linguistics – a vanishingly rare skill set. That said, the implementation of a natural language system for querying of journal articles to extract functional gene annotation is clearly a superior strategy. One caveat however, is that good annotation will only be found for extensively studied model organisms and economically important crops. However, a combination of full text searching and gene relationship characterization by orthology can close the annotation gap in many species.

This means that using typical annotation strategies for most plant species, and especially for orphan species, the unannotatable gene component is well over 50% of a plant's gene complement and functional gene annotation is still lagging well behind advances in genome assembly. Moreover, the reliance on GO for genome annotation is a practice that needs to be supplemented by other methodologies, as even in the best studied of all plant species, *A. thaliana*, only 66% of all genes can be annotated via GO. This is the maximum annotatable gene space in any plant, and for orphan crops, it will be much lower.

It should be noted that the majority of human genes have been manually annotated by the Sanger Centre's HAVANA group (https://www.sanger.ac.uk/science/projects/manual-annotation) and have not been annotated by computational means. Human annotation remains the gold standard for gene functional annotation, but natural language processing applications make automated approximations of human annotation strategies a reality.

The Sugarcane, Miscanthus and Fonio Blanc Lignin Pathways

In a pathway and transcriptomic context, sugarcane (*Saccharum* hybrid) is an orphan crop, while *D. exilis* (known as fonio or fonio blanc in Senegal) is a true orphan small grain crop, with only a single high-depth dataset (SRR3938613) deposited in NCBI's Sequence Read Archive (SRA). Miscanthus, in contrast, has near-complete genome assembly, but the majority of its lignin pathway genes remain unidentified and unannotated. We illustrate the principles of transcript and gene assembly, gene annotation, structure prediction conformation of gene function, novel gene discovery by natural language text processing, and pathway construction using sugarcane, miscanthus and *D. exilis* as our models.

Both KEGG and BioCYC have reference lignin biosynthesis pathways. In KEGG, the lignin biosynthetic pathway comes under the larger phenylpropanoid biosynthesis pathway. As *A. thaliana* is the best-annotated and studied plant genome, this was taken as the reference pathway. The pathway was manually reduced to the monolignol biosynthesis component only. This reduced KEGG pathway was compared with the BioCYC pathway and differences were resolved. As the lignin biosynthesis pathway of *Populus triocarpa* (black cottonwood) is the most well studied, gene names were taken from the *Populus* pathway. Arabidopsis gene identifiers were associated with the key enzymes to generate the following consensus pathway (Fig. 7.9).

As assembling genes/transcripts based on a reference is far more efficient than *de novo* assembly, the genes identified in *A. thaliana* were mapped to the genome reference of the most closely related species (*S. bicolor* for sugarcane and *Setaria viridis* for *D. exilis*) using Ensembl orthology at the gene level. Table 7.3 contains the cognate genes identified.

Use of the bait and assemble methodology that we had previously employed for assembling and characterizing sugarcane herbicide targets and detoxification proteins (Lloyd Evans and Joshi, 2017) yielded complete gene and transcript assemblies for the sugarcane and white fonio cognates of all the genes in the lignin pathway as above, apart from SAD, which could not even be assembled with protein targets using Diamond and appears to be a dicot-specific enzyme.

All the assembled genes yielded the expected transcripts and the transcripts translated cleanly. As the gene products were enzymes, their protein translations were submitted to the Phyre[2] server for homology modelling. In all cases, the structures obtained were for the correct enzyme, demonstrating that our methodology had identified and assembled the cognate orthologue of the Arabidopsis/maize enzyme within the pathway. A selection of structural models for the assembled enzymes is given in Fig. 7.10.

Natural language processing analyses, searching for lignin pathway genes focused on

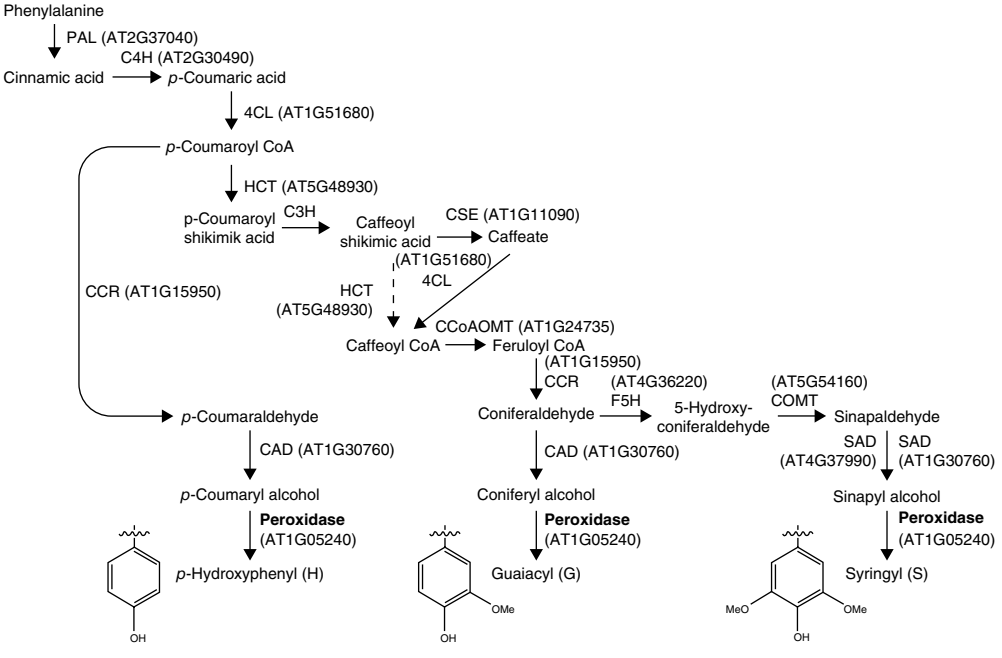


Fig. 7.9. Consensus lignin biosynthesis pathway for *Arabidopsis thaliana.* A schematic of the main steps in lignin formation for *A. thaliana* generated as a consensus from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioCYC pathways. The main genes are: PAL – phenylalanine ammonia-lyase; C4H – cinnamate 4-hydroxylase; 4CL – 4-coumarate:CoA ligase; HCT – hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase; C3H – coumarate 3-hydroxylase; CCR – cinnamoyl-CoA reductase; CSE – caffeoyl shikimate esterase; CCoAOMT – caffeoyl coenzyme A O-methyltransferase; F5H – ferulate 5-hydroxylase; CAD – cinnamoyl alcohol dehydrogenase; SAD – sinapyl alcohol dehydrogenase. The final monolignol products of the pathway, *p*-hydroxyphenyl (H-lignin), guaiacyl (G-lignin) and syringyl (S-lignin) are shown as structural models.

maize and the Andropogoneae, revealed maize orthologues for the two lignin genes with no grass orthologues in Ensembl (Table 7.3). Also, an additional monocot-specific lignin biosynthesis gene (bifunctional PTAL) was identified along with genes involved in lignin polymerization.

The extended sugarcane lignin pathway (which is also the pathway in maize, sorghum, miscanthus and white fonio) is given in Fig. 7.11. This includes the grass-specific bifunctional PTAL enzyme, which can convert either phenylalanine or tyrosine into lignin precursors as well as the biosynthesis of ferulate (Barros *et al.*, 2016), which is a unique lignin cross-linker in grasses. The pathway is also extended to incorporate UDP-glycosyltransferase 72E2 and E3 along with β-glucosidase and peroxidase/laccase. This makes our extended lignin pathway in sugarcane, miscanthus and white fonio the most comprehensively mapped pathway in any grass.

Transcript sequence confirmation by sequencing

Using the transcript sequences identified by sequence assembly as templates, primers were designed that spanned the 5′ UTR and start codon as well as the stop codon and 3′ UTR. These were employed to amplify the sugarcane and *D. exilis* transcripts from the lignin pathway. Amplicons were eluted through an HPLC column and were sequenced with MinION 9.5 chemistry, with samples prepared using the PCR Sequencing Kit (SQK-PSK004). Full-length reads with 99% accuracy were obtained in a 6-h sequencing run. Reads were assembled and error corrected with the Canu assembler (Koren *et al.*, 2017) and revealed that our assembly protocol had identified and assembled the correct transcript in each case. Sequences for transcripts in the sugarcane lignin pathway can be obtained from

Table 7.3. List of lignin biosynthesis genes identified in *Arabidopsis thaliana* and their orthologues in *Sorghum bicolor* and *Setaria italica*.

Gene name	Gene code	*Arabidopsis thaliana* gene	*Sorghum bicolor* orthologue	*Setaria italica* orthologue
4-Coumarate-CoA ligase	4CL	At1G51680	SORBI_3010G214900	SETIT_006172mg
p-Coumaroyl quinate/shikimate 3′-hydroxylase	C3′H/CYP98A3	AT2G40890	SORBI_3009G181800	SETIT_021789mg
Cinnamate-4-hydroxylase	C4H/CYP73A5	At2G30490	SORBI_3003G337400	SETIT_001126mg
Cinnamyl alcohol dehydrogenase	CAD	AT1G72680	SORBI_3010G072000	SETIT_039271mg
Phenylalanine ammonia-lyase	PAL	At2G37040	SORBI_3006G148800[a]	SETIT_012256mg[a]
Cinnamoyl-CoA reductase 1	CCR1	At1G15950	SORBI_3010G066000	SETIT_006847mg
Caffeoyl-CoA O-methyltransferase	CCoAOMT/OMT1	At1G24735	SORBI_3010G052200[b]	–
Hydroxycinnamoyl-CoA shikimate transferase	HCT	At5G48930	SORBI_3004G212300	SETIT_011974mg
Caffeoyl shikimate esterase	CSE	At1G11090	SORBI_3002G341600	SETIT_030452mg
Ferulic acid 5-hydroxylase	F5H/CYP84A1	At4G36220	SORBI_3002G029500	SETIT_035174mg
Flavone 3′-O-methyltransferase	COMT/OMT	At5G54160	SORBI_3007G04730[c]	SETIT_014900mg[c]
Sinapyl alcohol dehydrogenase	SAD	At4G37990	–	–
Peroxidase	PER	AT1G05260	SORBI_3010G232500	SETIT_006871mg

This table lists the main enzymes in the lignin biosynthesis pathway, with their gene codes, gene identifiers in *Arabidopsis thaliana* and cognate genes in the *Sorghum bicolor* and *Setaria italica* genome references. Superscripts within the table indicate the following: [a]No orthologue was identified from the Arabidopsis genome, the gene was identified from *Zea mays* by text searching and orthologues were identified against Zm00001d003016. [b]No close orthologue was identified from the Arabidopsis gene, grass CCoAOMT orthologues were identified from *Zea mays* by natural language searching and the Sorghum orthologue was identified against Zm00001d036293, but no orthologue was identified for *Setaria italic*. [c]No close orthologue could be identified from the Arabidopsis gene. Grass cognates were identified from the brown midrib mutation in the *Zea mays* gene Zm00001d049541 using full text searching.

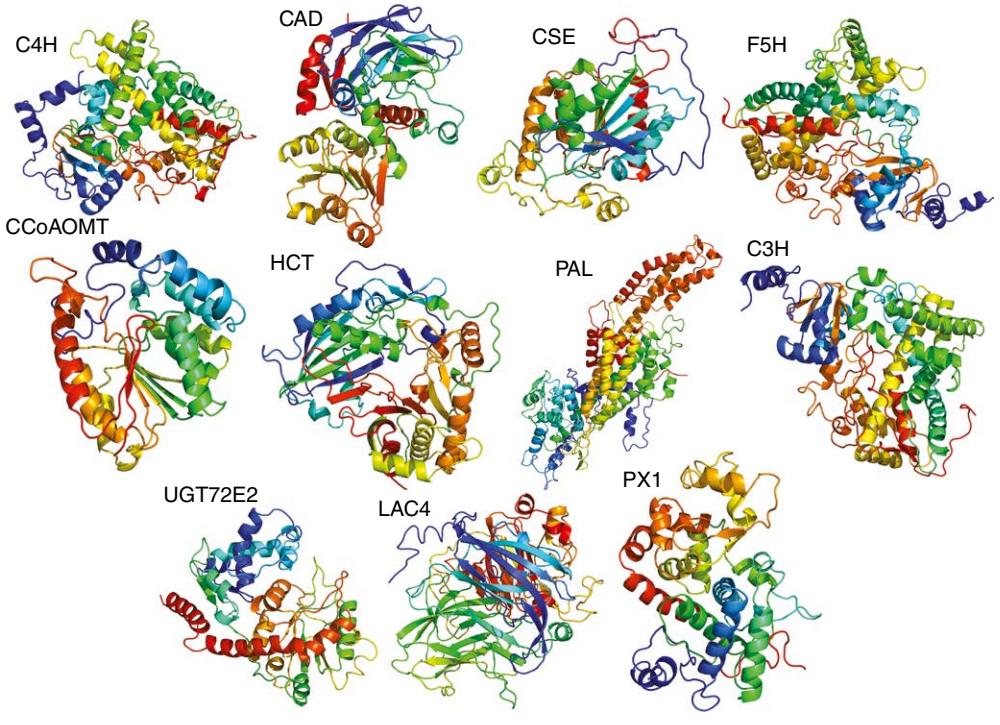


Fig. 7.10. Structural models for a select subset of lignin biosynthesis genes assembled in sugarcane (*Saccharum* hybrid), miscanthus and white fonio (*Digitaria exilis*). Images shown are structural models for sugarcane proteins, however, both miscanthus and *D. exilis* models were topologically nearly identical (less than 0.3Å RMSD deviation). Models presented are: C4H – cinnamate 4-hydroxylase; CAD – cinnamoyl alcohol dehydrogenase; CSE – caffeoyl shikimate esterase; F5H – ferulate 5-hydroxylase; CCoAOMT – caffeoyl coenzyme A O-methyltransferase; HCT – hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase; PAL – phenylalanine ammonia-lyase; C3H – coumarate 3-hydroxylase; UGT72E2 – UDP glycosyltransferase 72E2; LAC4 – laccase 4 and PX1 – peroxidase 1.

ENA (European Nucleotide Archive) via the project identifier PRJEB29703 for sugarcane, PRJEB30269 for *M. sinensis* cv. Andante and PRJEB29768 and for *D. exilis*.

Taken together, our approach of orthologue discovery, template-based assembly, confirmation of gene function by structural modelling, novel gene discovery and annotation by natural language processing has demonstrated that high accuracy and efficacy gene assembly and annotation can be performed in orphan crops and species. Moreover, all these studies can be performed on a high-end laptop.

Regulation of the lignin pathway

Along with extending the lignin pathway to grass-specific genes and secondary biosynthetic routes, we have also used full text searching to identify genes that regulate lignin biosynthesis. These genes include:

- **ZmSWN6 NAC** transcription factor, a master regulator of secondary wall assembly and lignification (Barrière *et al.*, 2018).
- **ZmMYB019** belongs to the MYB G13 group, of which several members have been shown to be involved in secondary wall biosynthesis and lignification (Barrière *et al.*, 2018).
- **ZmMYB42** enriches H- and G-lignin at the expense of S-lignin and is involved in controlling the phenylpropanoid biosynthetic pathway (Sonbol *et al.*, 2009).
- **KN1** *Knotted1-like-homeobox* (*KNOX*) genes regulate at least two steps in the lignin biosynthesis pathway and effectively

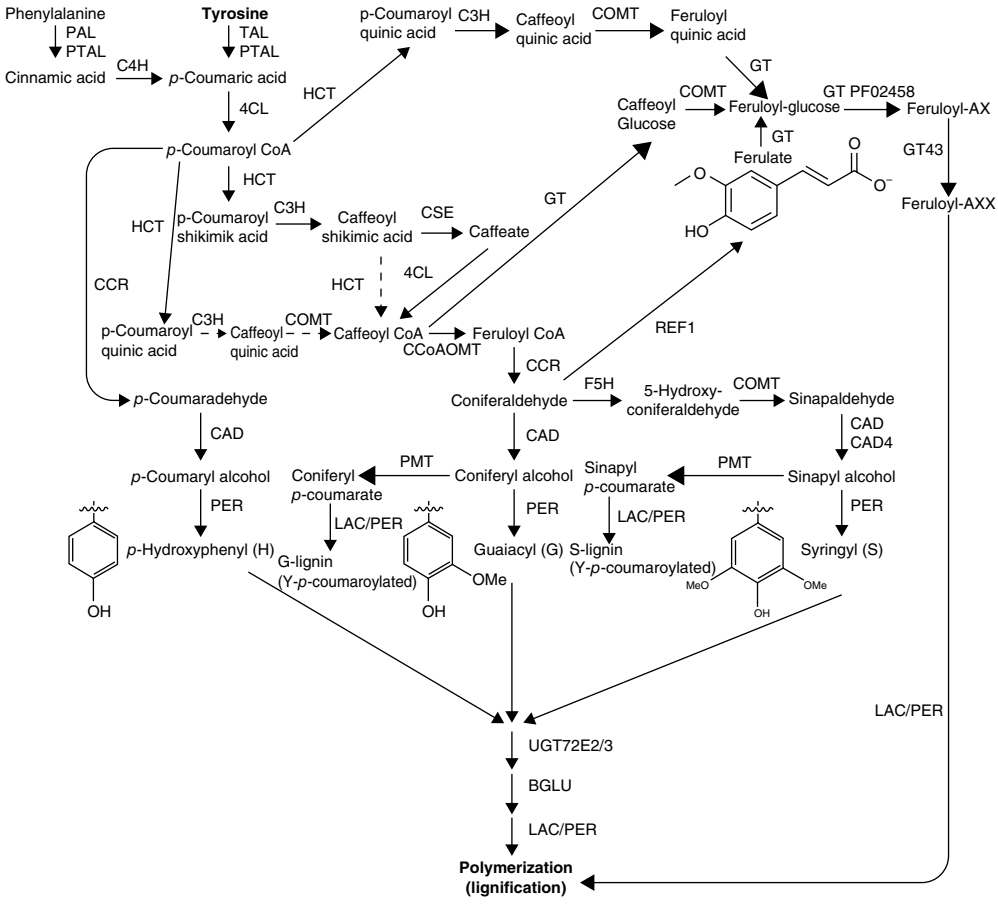


Fig. 7.11. Extended lignin biosynthesis and lignification pathway in sugarcane, miscanthus and white fonio. Schematic of the lignin biosynthesis and lignification pathway is sugarcane (*Saccharum* hybrid), miscanthus (*Miscanthus sinensis* cv. Andante) and white fonio (*Digitaria exilis*). This pathway has been extended with the grass-specific PTAL (phenylalanine/tyrosine bifunctional ammonia-lyase) along with the feruloyl biosynthesis and cross-linking pathway. The main genes are: PAL – phenylalanine ammonia-lyase; C4H – cinnamate 4-hydroxylase; 4CL – 4-coumarate:CoA ligase; HCT – hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase; C3H – coumarate 3-hydroxylase; CCR – cinnamoyl-CoA reductase; CSE – caffeoyl shikimate esterase; CCoAOMT – caffeoyl coenzyme A O-methyltransferase; F5H – ferulate 5-hydroxylase; CAD – cinnamoyl alcohol dehydrogenase; SAD – sinapyl alcohol dehydrogenase; TAL – tyrosine ammonia-lyase; COMT – caffeic acid O-methyltransferase; GT – glucosyl transferase; PER – peroxidase; LAC – laccase; REF1 – an aldehyde dehydrogenase known as *REDUCED EPIDERMAL* FLUORESCENCE1 in Arabidopsis. For the formation of feruloyl-Ax and feruloyl-AXX the specific glucosyl transferase isoform involved in the reaction is named. Structures for ferulate and the three monolignols, *p*-hydroxyphenyl (H-lignin), guaiacyl (G-lignin) and syringyl (S-lignin) are given.

downregulate lignification in both grasses and dicots (Townsley *et al.*, 2013).

- **MYB15** is a regulator of defence-induced lignification and basal immunity (Chezem *et al.*, 2017).
- **MYB4** is a self-induced lignin biosynthetic repressor, which downregulates lignin production and upregulates sucrose biosynthesis. Overexpression of MYB4 in switchgrass (*P. virgatum*) can increase sugar release efficiency as much as threefold (Shen *et al.*, 2013).
- **NST1/NST2/NST3** act as primary regulators of secondary cell wall formation,

where NTS3 acts directly to upregulate F5H; NST1 upregulates MYB46; and the combination of MYB83 and NTS2 upregulates cellulose and xylan production (Mitsuda *et al.*, 2007; Zhao and Dixon, 2011).

- **SND1** acts redundantly to NST1 as a primary regulator of secondary cell wall formation (Zhong *et al.*, 2007). Loss of NST1 expression leads to a reduction in F5H activity and a significant reduction in S-liginin (Zhao *et al.*, 2010).
- **BREVIPEDICELLUS** (BP) is an Arabidopsis KNOX gene that is involved in internode patterning and is a repressor of lignification genes. It is downregulated by auxins and induces increased expression of MYB4 (Mele *et al.*, 2003).
- **WRKY2** is a transcription factor that plays a role in plant response to phytopathogens. It results in an upregulation of CH4, increasing G-lignin and reducing S-lignin production (Guillaumie *et al.*, 2010).
- **MYB46 and MYB83** act as master regulators to upregulate cellulose and xylan generation and upregulate MYB4 expression (Zhao and Dixon, 2011).
- **MYB32** is upregulated by auxins and acts as a repressor of lignin biosynthesis (Zhao and Dixon, 2011).
- **MYB85, MYB63** and **MYB58** are regulators of lignin biosynthesis. They are upregulated by MYB46 and MYB83 and act via AC elements to induce secondary wall thickening by lignification (Zhou *et al.*, 2009).
- **MYB26** induces lignification and secondary wall thickening by upregulating the expression of NST1 and NST2 (Xie *et al.*, 2018).
- **MYB4, MYB32 and MYB7** are induced by MYB46 and act to downregulate the lignin biosynthetic pathway via AC elements (Xie *et al.*, 2018).
- **MYB75** acts as a repressor of both cellulose/xylan biosynthesis and lignin biosynthesis. However, it upregulates the biosynthesis of anthocyanins (Zhao and Dixon, 2011).

Our current best knowledge about the regulation of lignification (based primarily on data from maize and Arabidopsis) is summarized in Fig. 7.12. This is another kind of pathway, a regulatory network that describes the way genes and entire pathway elements are regulated.

Though chalcone synthase was first identified as a potential silencing factor in maize that increases lignin content, it impedes the incorporation of tricin into lignin (Eloy *et al.*, 2017). It is now known that it is a key enzyme in tricin biosynthesis, and this has led to a major extension to the lignification (lignin polymerization) pathway.

Extending the lignin pathway: an improved understanding of lignification

The discovery that tricin is required to initiate lignin chains in grasses (Eloy *et al.*, 2017) has revealed a major and previously hidden pathway in lignification. Tricin biosynthesis shares a common key precursor with the lignin biosynthetic pathway, *p*-coumaroyl CoA. The pathway was initially uncovered in maize, but we demonstrate cognate pathways in sugarcane, miscanthus and *D. exilis*. The pathway is summarized in Fig. 7.13.

Applications – quantitative PCR and expression microarrays

The identification of a set of genes/transcripts of known function or belonging to a mapped pathway leads to an immediate application within molecular breeding – the use of the annotated sequence for transcriptomic analysis and functional gene screening. Once a functional gene has been identified, additional text mining can be performed to examine the effect of up- and down-regulation of that gene. For example, upregulation of S-lignin significantly increases the digestibility of plant cell walls both for herbivores and bacterial-based biorefineries. A key enzyme in the lignification pathway for S-lignin generation is F5H and upregulation of F5H in rice (*O. sativa* L.) increases S-lignin while downregulating C-lignin (Takeda *et al.*, 2017). Downregulation of CAD1 in *Brachypodium distachyon* resulted in an increase in S-lignin accumulation in the plant's cell walls (Bouvier d'Yvoire *et al.*, 2013) while concomitantly resulting in an increase in saccharification. Combined downregulation of CCR and COMT or CCR and CAD also results in an upregulation of S-lignin and a

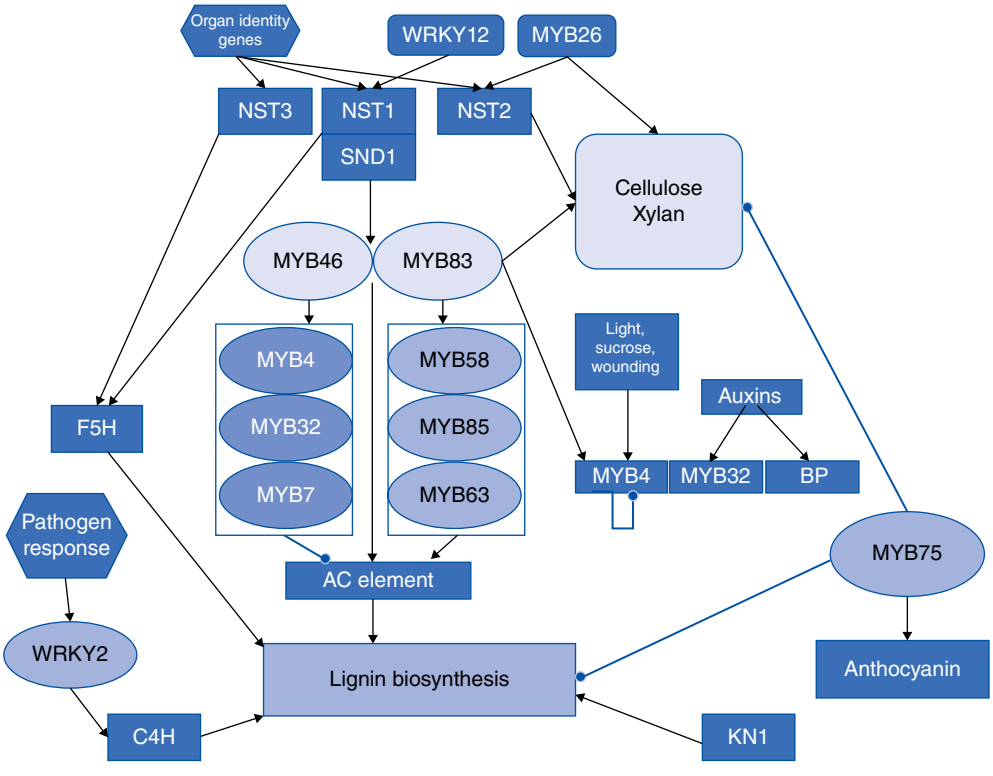


Fig. 7.12. Schematic summarizing our current best knowledge about the regulatory networks influencing lignin biosynthesis. In the diagram, lines ending in arrows show pathways of activation and lines ending in circles represent inhibitory pathways. The AC element is a common cis-regulatory element present in the majority of lignin-pathway genes. Functions of all the regulatory proteins are detailed in the main text.

down-regulation of G-lignin (Baucher *et al.* 2003). In grasses, downregulation of *Brachypodium distachyon* PTAL-1 indicated that the enzyme can provide nearly half of the total lignin deposited in *Brachypodium*, with a preference for S-lignin and wall-bound coumarate biosynthesis (Barros *et al.*, 2016). In contrast, up-regulation of REF1 (ALDH) leads to an up-regulation of ferulate but a relative down-regulation of both G and S lignins. However, increased ferulate induces more G-lignin cross-linking, which may be associated with increased resistance to stem borers (Santiago *et al.*, 2016). Increased H-lignin and increased *p*-coumaroyl-COA also seem to be associated with borer resistance, indicating that increasing CAD expression, while concurrently reducing CCoAOMT expression, could be a pathway to stem borer-resistance traits in grasses.

As the sequences of all the above-referenced genes are known and confirmed by sequencing in sugarcane, miscanthus and white fonio, it is possible to develop PCR primers for all these genes. These primers can be utilized in reverse transcription quantitative PCR (RT-qPCR) (Taylor *et al.*, 2010) with relative expression levels determined against three reference genes for each tissue studied.

Where there are larger numbers of full-length transcripts available, e.g. for the complete lignin pathway and saccharogenesis pathways in sugarcane, it is possible to use these as the basis for expression microarray production (as detailed in Popp *et al.*, 2007). These microarrays can then be used for large-scale tissue and whole-germplasm collection analyses to identify expression variations within the germplasm. From these results, parents can be chosen for selective breeding of desirable lignification traits – or any other trait. Figure 7.14 contains a schematic summary of how a cDNA-based microarray (gene chip) works for gene expression profiling of a

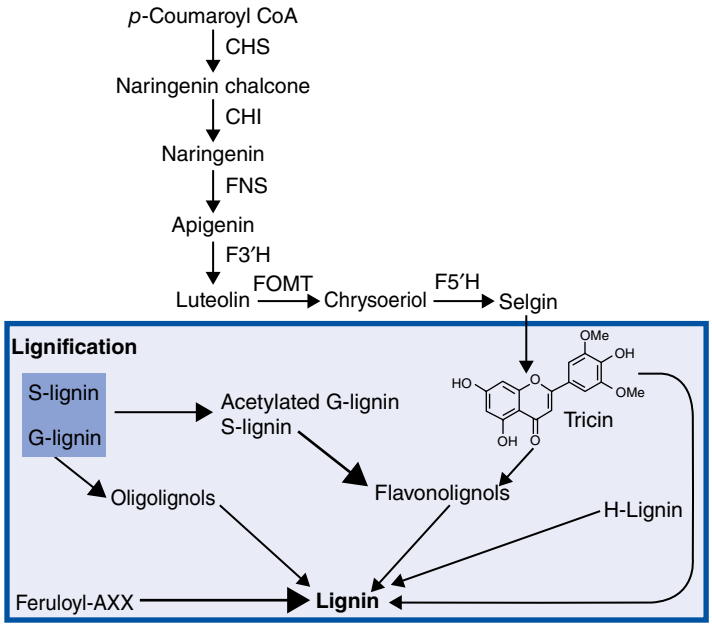


Fig. 7.13. The tricin biosynthetic pathway in grasses. The tricin biosynthetic pathway links out from the core lignin pathway via its primary input, *p*-coumaroyl CoA, and links back to the pathway during lignin polymerization, as tricin is the main seed for lignification in grasses. The main enzymes in the pathway are: CHS – chalcone synthase; CHI – chalcone isomerase; FNS – flavone synthase; F3′H – flavonoid 3′-hydroxylase; FOMT – flavonoid O-methyltransferase and F5′H – ferulate 5-hydroxylase. Tricin is shown as a structural model.

sample against a reference. As the DNA library is tagged with a fluorescent label and only those sequences that bind to the probes on the chip fluoresce, the level of fluorescence (luminosity) is proportional to the amount of bound DNA. This means that large numbers of genes can be analysed very rapidly for a given phenotype; the major limitation being that these chips are single-use only and a new chip is required for each experiment (each pair of genotypes to be analysed). This makes RT-qPCR more cost-effective for analysing a small number of genes across multiple genotypes, particularly as using the same reference genes, multiple genotypes can be directly compared within a single experiment.

Multiple pathways, along with the sequences for each transcript within the pathway, can be constructed with the techniques detailed above. From the transcripts assembled (including alternate transcripts), unique 90-mer sequences can be extracted and these can be used as probes for expression chips. The more pathway members are assembled, the broader the range of biological questions that can be answered and the more cost-effective is the use of a microarray chip. To this end, we have used literature mining and pathway reconstruction to assemble the saccharogenesis pathways and sugar transport systems in sugarcane. We have also identified and assembled more than 320 stress-response genes in sugarcane. In total, we have more than 8000 assembled sugarcane genes and 12,000 transcripts associated with sugarcane traits for the development of a sugarcane expression chip. All these genes, along with associated functions and phenotypes, have been identified and assembled using precisely the techniques detailed in this chapter.

For sugarcane, we have gone from having only limited genomic data to being able to generate an expression microarray that allows for the analysis of thousands of genes with potential agronomic importance. Even more spectacularly, we have been able to analyse the genome of *D. exilis*, a true orphan crop, in parallel, and this would enable us to generate an equivalent expression microarray for this species, which

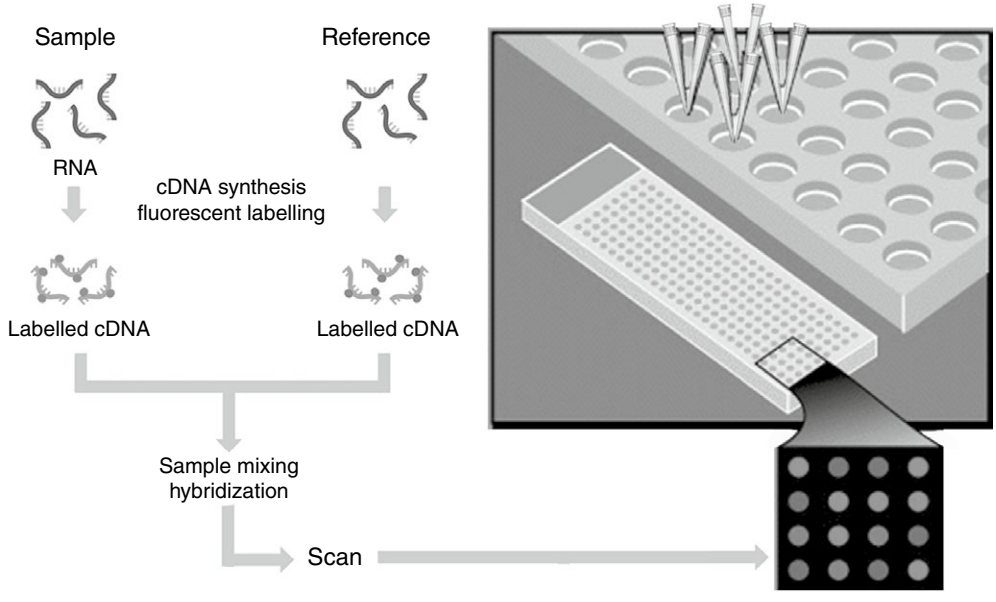


Fig. 7.14. Schematic image of a microarray experiment. The image shows a schematic view of how a microarray experiment works. Each well or segment of the chip is labelled with a 90-mer probe specific for a certain gene or transcript. The cDNA library from both reference and sample genotypes are labelled with different fluorophores before being bound to the primers on the chip. After washing, only those cDNAs bound to the probes are left. The 'chip' or microarray is scanned and the difference in fluorescence for the two fluorophores reveals the relative expression of a given gene/transcript between the sample and the reference.

has no genomic sequence available. Our techniques have also allowed for the functional annotation and sequence confirmation of miscanthus transcripts.

Moreover, for transcript assembly, we have incorporated PacBio transcriptomic data into our assembly pipeline based on Diamond protein mapping. This appears to yield much better and more accurate full-length transcripts, with the transcripts translating cleanly.

Conclusion

In this chapter, we have outlined approaches for gene assembly and gene annotation of orphan crops that allow for sequence assembly even if no closely related sequence is available. We have demonstrated the utility of full text mining for gene annotation and pathway discovery and shown that it can be substantially superior to the current 'best in class' annotation methodology, mapping to GO terms. Using pathway mapping, we have achieved significant advances in the understanding of lignification both at the enzymatic and control levels. We have shown that text mining can be the key to understanding and extending existing pathways.

Using *D. exilis* as an exemplar, we have shown that the systems designed for sugarcane are applicable to any orphan crop. Moreover, all the bioinformatics platforms and techniques discussed can be deployed on just a high-end laptop or desktop running Linux, putting the techniques within the reach of most research groups.

References

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., *et al*. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications* 7, 11706.

Amundson, R. and Lauder, G.V. (1994) Function without purpose. *Biology and Philosophy* 9(4), 443–469.
Anderson, E., Arundale, R., Maughan, M., Oladeinde, A., Wycislo, A., *et al*. (2011) Growth and agronomy of *Miscanthus* x *giganteus* for biomass production. *Biofuels* 2(1), 71–87.
Andorf, C.M., Cannon, E.K., Portwood, J.L., Gardiner, J.M., Harper, L.C., *et al*. (2015) MaizeGDB update: New tools, data and interface for the maize model organism database. *Nucleic Acids Research* 44(D1), D1195–D1201.
Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium*, American Medical Informatics Association, p.17.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., *et al*. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5), 455–477.
Barrière, Y., Guillaumie, S., Denoue, D., Pichon, M., Goffner, D. *et al*. (2018) Investigating the unusually high cell wall digestibility of the old INRA early flint F4 maize inbred line. *Maydica* 62(3), 21.
Barros, J., Serrani-Yarce, J.C., Chen, F., Baxter, D., Venables, B.J., *et al*. (2016) Role of bifunctional ammonia-lyase in grass cell wall biosynthesis. *Nature Plants* 2(6), 16050.
Baucher, M., Halpin, C., Petit-Conil, M. and Boerjan, W. (2003) Lignin: Genetic engineering and impact on pulping. *Critical Reviews in Biochemistry and Molecular Biology* 38(4), 305–350.
Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., *et al*. (2015) The Arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome. *Genesis* 53(8), 474–485.
Berman, H., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide protein data bank. *Nature Structural and Molecular Biology* 10(12), 980.
Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc. Sebastopol, California.
Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., *et al*. (2014) Plant genome sequencing – applications for crop improvement. *Current Opinion in Biotechnology* 26, 31–37.
Bolser, D., Staines, D.M., Pritchard, E. and Kersey, P. (2016) Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards, D. (ed.) *Plant Bioinformatics*. Springer, New York, pp. 115–140.
Bond, D.M., Albert, N.W., Lee, R.H., Gillard, G.B., Brown, C.M., *et al*. (2016) Infiltration-RNAseq: Transcriptome profiling of Agrobacterium-mediated infiltration of transcription factors to discover gene function and expression networks in plants. *Plant Methods* 12(1), 41.
Bouvier d'Yvoire, M., Bouchabke-Coussa, O., Voorend, W., Antelme, S., Cézard, L., *et al*. (2013) Disrupting the cinnamyl alcohol dehydrogenase 1 gene (Bd CAD 1) leads to altered lignification and improved saccharification in *Brachypodium distachyon*. *The Plant Journal* 73(3), 496–508.
Brückner, A., Polge, C., Lentze, N., Auerbach, D. and Schlattner, U. (2009) Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences* 10(6), 2763–2788.
Buchfink, B., Xie, C. and Huson, D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12(1), 59.
Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., *et al*. (2008) AmiGO: Online access to ontology and annotation data. *Bioinformatics* 25(2), 288–289.
Carvalho, D.S., Schnable, J.C. and Almeida, A.M.R. (2018) Integrating phylogenetic and network approaches to study gene family evolution: The case of the AGAMOUS family of floral genes. *Evolutionary Bioinformatics* 14, 1176934318764683.
Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., *et al*. (2007) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 36(suppl. 1), D623–D631.
Chen, L., Auh, C.K., Dowling, P., Bell, J., Lehmann, D., *et al*. (2004) Transgenic down-regulation of caffeic acid O-methyltransferase (COMT) led to improved digestibility in tall fescue (*Festuca arundinacea*). *Functional Plant Biology* 31(3), 235–245.
Chezem, W.R., Memon, A., Li, F.S., Weng, J.K. and Clay, N.K. (2017) SG2-type R2R3-MYB transcription factor MYB15 controls defense-induced lignification and basal immunity in Arabidopsis. *The Plant Cell* 29(8), 1907–1926.
Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., *et al*. (2012) The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology* 54(2), e1.
Coordinators, NCBI Resource, Acland, A., Agarwala, R., Barrett, T., Beck, J., *et al*. (2014) Database resources of the national center for biotechnology information. *Nucleic Acids Research* 42(Database issue), D7.

Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. (2013) Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology* 9(2), e1002854.

Daniels, J. and Roach, B.T. (1987) Taxonomy and evolution. In: Heinz, D.J. (ed.) *Sugarcane Improvement Through Breeding*, Volume 11. Elsevier, Amsterdam, pp. 7–84.

Dash, S., Campbell, J.D., Cannon, E.K., Cleary, A.M., Huang, W., *et al*. (2015) Legume information system (LegumeInfo.org): A key component of a set of federated data resources for the legume family. *Nucleic Acids Research* 44(D1), D1181–D1188.

Dash, S., Cannon, E.K., Kalberer, S.R., Farmer, A.D. and Cannon, S.B. (2016) PeanutBase and other bioinformatic resources for peanut. In: Stalker, H.T. and Wilson, R.F. (eds) *Peanuts: Genetics, Processing, and Utilization*. Elsevier, Amsterdam, pp. 241–252.

D'Hont, A. (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenetic and Genome Research* 109(1–3), 27–33.

Eloy, N.B., Voorend, W., Lan, W., Saleme, M.D.L.S., Cesarino, I., *et al*. (2017) Silencing chalcone synthase in maize impedes the incorporation of tricin into lignin and increases lignin content. *Plant Physiology* 173(2), 998–1016.

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A. and Pappas, G. (2008) Comparison of PubMed, Scopus, web of science, and Google scholar: Strengths and weaknesses. *The FASEB Journal* 22(2), 338–342.

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., *et al*. (2014) The Sol Genomics Network (SGN) – from genotype to phenotype to breeding. *Nucleic Acids Research* 43(D1), D1036–D1041.

Garí, J.A. (2002) Review of the African millet diversity. *International Workshop on Fonio, Food Security and Livelihood among the RURAL Poor in West Africa*, Bamako, Mali, IPGRI/IFAD, pp. 19–22.

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., *et al*. (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature Communications* 9(1), 2638.

Głowacka, K., Ahmed, A., Sharma, S., Abbott, T., Comstock, J.C., *et al*. (2016) Can chilling tolerance of C4 photosynthesis in *Miscanthus* be transferred to sugarcane? *GCB Bioenergy* 8(2), 407–418.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., *et al*. (2011) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* 40(D1), D1178–D1186.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7), 644.

Grivet, L. and Arruda, P. (2002) Sugarcane genomics: Depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology* 5(2), 122–127.

Guillaumie, S., Mzid, R., Méchin, V., Léon, C., Hichri, I., *et al*. (2010) The grapevine transcription factor WRKY2 influences the lignin pathway and xylem development in tobacco. *Plant Molecular Biology* 72(1–2), 215.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., *et al*. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research* 44(D1), D1214–D1219.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., *et al*. (2016) Ensembl comparative genomics resources. *Database* 2016, bav096. DOI: 10.1093/database/bav096.

Hirsch, C.N. and Robin Buell, C. (2013) Tapping the promise of genomics in species with complex, nonmodel genomes. *Annual Review of Plant Biology* 64, 89–110.

Hoang, N.V., Furtado, A., Botha, F.C., Simmons, B.A. and Henry, R.J. (2015) Potential for genetic improvement of sugarcane as a source of biomass for biofuels. *Frontiers in Bioengineering and Biotechnology* 3, 182.

Hodkinson, T.R., Chase, M.W. and Renvoize, S.A. (2002a) Characterization of a genetic resource collection for *Miscanthus* (Saccharinae, Andropogoneae, Poaceae) using AFLP and ISSR PCR. *Annals of Botany* 89(5), 627–636.

Hodkinson, T.R., Chase, M.W., Takahashi, C., Leitch, I.J., Bennett, M.D., *et al*. (2002b) The use of DNA sequencing (ITS and trnL-F), AFLP, and fluorescent in situ hybridization to study allopolyploid *Miscanthus* (Poaceae). *American Journal of Botany* 89(2), 279–286.

Hoehndorf, R., Schofield, P.N. and Gkoutos, G.V. (2015) The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics* 16(6), 1069–1080.

Hofstrand, D. (2009) Brazil's ethanol industry. Available at: https://www.extension.iastate.edu/agdm/articles/hof/HofFeb09.html (accessed 23 November 2018).

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., *et al*. (2016) Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PLoS ONE* 11(1), e0146062.

Irvine, J.E. (1999) Saccharum species as horticultural classes. *Theoretical and Applied Genetics* 98(2), 186–194.

Jeżowski, S., Mos, M., Buckby, S., Cerazy-Waliszewska, J., Owczarzak, W., *et al*. (2017) Establishment, growth, and yield potential of the perennial grass *Miscanthus* × *giganteus* on degraded coal mine soils. *Frontiers in Plant Science* 8, 726.

Jiao, W.B. and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* 36, 64–70.

Jideani, I.A. (1999) Traditional and possible technological uses of *Digitaria exilis* (acha) and *Digitaria iburua* (iburu): A review. *Plant Foods for Human Nutrition* 54(4), 363–374.

Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M.A., *et al*. (2009) NCBO annotator: Semantic annotation of biomedical data. *International Semantic Web Conference, Poster and Demo session*, October 2009, Vol. 110.

Jung, S., Ficklin, S.P., Lee, T., Cheng, C.H., Blenda, A., *et al*. (2013) The genome database for Rosaceae (GDR): Year 10 update. *Nucleic Acids Research* 42(D1), D1237–D1244.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28(1), 27–30.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10(6), 845.

Kim, J.A., Roy, N.S., Lee, I.H., Choi, A.Y., Choi, B.S., *et al*. (2019) Genome-wide transcriptome profiling of the medicinal plant *Zanthoxylum planispinum* using a single-molecule direct RNA sequencing approach. *Genomics* 111, 973–979.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. *et al*. (2017) Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* 27, 722–736.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., *et al*. (2014) Araport: The Arabidopsis information portal. *Nucleic Acids Research* 43(D1), D1003–D1009.

Kumatoo (2009) Sanoussi Diakite. Available at: http://africaninnovation.org/?persons=sanoussi-diakite (accessed 18 November 2018).

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., *et al*. (2011) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research* 40(D1), D1202–D1210.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13(9), 2178–2189.

Li, X., Ximenes, E., Kim, Y., Slininger, M., Meilan, R., *et al*. (2010) Lignin monomer composition affects Arabidopsis cell-wall degradability after liquid hot water pretreatment. *Biotechnology for Biofuels* 3(1), 27.

Liu, Q., Luo, L. and Zheng, L. (2018) Lignins: Biosynthesis and biological functions in plants. *International Journal of Molecular Sciences* 19(2), 335.

Lloyd Evans, D. and Joshi, S.V. (2016) Complete chloroplast genomes of *Saccharum spontaneum*, *Saccharum officinarum* and *Miscanthus floridulus* (Panicoideae: Andropogoneae) reveal the plastid view on sugarcane origins. *Systematics and Biodiversity* 14(6), 548–571.

Lloyd Evans, D. and Joshi, S.V. (2017) Herbicide targets and detoxification proteins in sugarcane: From gene assembly to structure modelling. *Genome* 60(7), 601–617.

Lloyd Evans D., Hlongwane, T.T., Joshi, S.V. and Riano-Pachón, D.M. (2019a) The sugarcane mitochondrial genome: Assembly, phylogenetics and transcriptomics. *PeerJ* 7, e7558. DOI: 10.7717/peerj.7558.

Lloyd Evans, D., Joshi, S. and Wang, J. (2019b) Whole chloroplast and gene locus phylogenies reveal the taxonomic placement and relationship of *Tripidium* (Panicoideae: Andropogoneae) to sugarcane. *BMC Evolutionary Biology* 19(1), 33.

Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics* 14(5), 265–279.

Lunter, G. and Goodson, M. (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21(6), 936–939.

McDermott, J., Bumgarner, R. and Samudrala, R. (2005) Functional annotation from predicted protein interaction networks. *Bioinformatics* 21(15), 3217–3226. DOI: 10.1093/bioinformatics/bti514.

Mele, G., Ori, N., Sato, Y. and Hake, S. (2003) The knotted1-like homeobox gene BREVIPEDICELLUS regulates cell differentiation by modulating metabolic pathways. *Genes & Development* 17(17), 2088–2093.

Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., *et al*. (2007) NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis. *The Plant Cell* 19(1), 270–280.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M. (2007) KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* 35, 182–185.

Müller, H.M., Kenny, E.E. and Sternberg, P.W. (2004) Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology* 2(11), e309.

Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., *et al*. (2016) The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* 45(D1), D712–D722.

Naithani, S., Preece, J., D'Eustachio, P., Gupta, P., Amarasinghe, V., *et al*. (2016) Plant Reactome: A resource for plant pathways and comparative analysis. *Nucleic Acids Research* 45(D1), D1029–D1039.

Nakasugi, K., Crowhurst, R., Bally, J. and Waterhouse, P. (2014) Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS ONE* 9(3), e91776.

NCBI CCDS (2018) The Consensus CDS (CCDS) Project, Version 22, 14 June 2018. Available at: https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi (accessed 24 November 2018).

Neander, K. (1991) Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58(2), 168–184.

Nikoloski, Z., Perez-Storey, R. and Sweetlove, L.J. (2015) Inference and prediction of metabolic network fluxes. *Plant Physiology* 169(3), 1443–1455.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., *et al*. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229), 551.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. and Salzberg, S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* 11(9), 1650.

Petryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A., Barrera, E., *et al*. (2015) Expression Atlas update – an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research* 44(D1), D746–D752.

Popp, M.P., Liu, L., Timmers, A., Esson, D.W., Shiroma, L., *et al*. (2007) Development of a microarray chip for gene expression in rabbit ocular research. *Molecular Vision* 13, 164.

Raven, P.H., Evert, R.F. and Eichhorn, S.E. (2005) Biology of Plants, 7th edn. W.H. Freeman, New York.

Rinaldi, R., Jastrzebski, R., Clough, M.T., Ralph, J., Kennema, M., *et al*. (2016) Paving the way for lignin valorisation: recent advances in bioengineering, biorefining and catalysis. *Angewandte Chemie International Edition* 55(29), 8164–8215.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., *et al*. (1999) Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Research* 27(1), 89–94. DOI: 10.1093/nar/27.1.89.

Santiago, R., Malvar, R.A., Barros-Rios, J., Samayoa, L.F. and Butrón, A. (2016) Hydroxycinnamate synthesis and association with Mediterranean corn borer resistance. *Journal of Agricultural and Food Chemistry* 64(3), 539–551.

Sauvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, T., Tatusova, T., *et al*. (2010) Gnomon – NCBI Eukaryotic Gene Prediction Tool. Available at: https://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf (accessed 21 November 2018).

Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein–protein interactions in yeast. *Nature Biotechnology* 18(12), 1257.

Shen, H., Mazarei, M., Hisano, H., Escamilla-Trevino, L., Fu, C., *et al*. (2013) A genomics approach to deciphering lignin biosynthesis in switchgrass. *The Plant Cell* 25(11), 4342–4361.

Shu, S., Goodstein, D., and Rokhsar, D. (2013) PERTRAN: Genome-guided RNA-seq Read Assembler. *Proceedings of the Cold Spring Harbor Lab Genome Informatics Conference*, Cold Spring Harbor, New York, 30 October–2 November 2013. Available at: https://www.osti.gov/servlets/purl/1241180 (accessed 21 November 2018).

Somerville, C., Youngs, H., Taylor, C., Davis, S.C. and Long, S.P. (2010) Feedstocks for lignocellulosic biofuels. *Science* 329(5993), 790–792.

Sonbol, F.M., Fornalé, S., Capellades, M., Encina, A., Tourino, S., *et al*. (2009) The maize ZmMYB42 represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in *Arabidopsis thaliana*. *Plant Molecular Biology* 70(3), 283.

Souza, G.M., Berges, H., Bocs, S., Casu, R., D'Hont, A., *et al*. (2011) The sugarcane genome challenge: Strategies for sequencing a highly complex genome. *Tropical Plant Biology* 4(3–4), 145–156.

Sreenivasan T.V., Ahloowalia B.S. and Heinz D.J. (2015) Cytogenetics. In: Heinz, D.J. (ed.) *Sugarcane Improvement Through Breeding*, Volume 11. Elsevier, Amsterdam, pp. 211–253.

Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J., *et al*. (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the USA* 105(19), 6959–6964.

Suslick, K.S. (1998) Sonochemistry. In: Kirk-Othmer (ed.) *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th edn. Wiley, New York, pp. 517–541.

Swindell, W.R. (2006) The association among gene expression responses to nine abiotic stress treatments in *Arabidopsis thaliana*. *Genetics* 174(4), 1811–1824.

Takeda, Y., Koshiba, T., Tobimatsu, Y., Suzuki, S., Murakami, S., *et al*. (2017) Regulation of CONIFERALDEHYDE 5-HYDROXYLASE expression to modulate cell wall lignin structure in rice. *Planta* 246(2), 337–349.

Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M. and Nguyen, M. (2010) A practical approach to RT-qPCR – publishing data that conform to the MIQE guidelines. *Methods* 50(4), S1–S5.

Tello-Ruiz, M.K., Stein, J., Wei, S., Youens-Clark, K., Jaiswal, P., *et al*. (2016) Gramene: A resource for comparative analysis of plants genomes and pathways. In: Edwards, D. (ed.) *Plant Bioinformatics*. Springer, New York, pp. 141–163.

Tennant, J.P., Waldner, F., Jacques, D.C., Masuzzo, P., Collister, L.B., *et al*. (2016) The academic, economic and societal impacts of Open Access: An evidence-based review. *F1000Research* 5, 632.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., *et al*. (2004) MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* 37(6), 914–939.

Tian, W. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology* 333(4), 863–882.

Townsley, B.T., Sinha, N.R. and Kang, J. (2013) KNOX1 genes regulate lignin deposition and composition in monocots and dicots. *Frontiers in Plant Science* 4, 121.

Tureček, F. (2002) Mass spectrometry in coupling with affinity capture–release and isotope-coded affinity tags for quantitative protein analysis. *Journal of Mass Spectrometry* 37(1), 1–14.

Ungaro, A., Pech, N., Martin, J.F., Mccairns, R.S., Mevy, J.P., *et al*. (2017) Challenges and advances for transcriptome assembly in non-model species. *PloS ONE* 12(9), e0185020.

UniProt Consortium, (2014) UniProt: A hub for protein information. *Nucleic Acids Research* 43(D1), D204–D212.

Wegrzyn, J.L., Lee, J.M., Tearse, B.R. and Neale, D.B. (2008) TreeGenes: A forest tree genome database. *International Journal of Plant Genomics* 412875. DOI: 10.1155/2008/412875.

Welker, C.A., Souza-Chies, T.T., Longhi-Wagner, H.M., Peichoto, M.C., McKain, M.R., *et al*. (2015a) Phylogenetic analysis of Saccharum s.l. (Poaceae; Andropogoneae), with emphasis on the circumscription of the South American species. *American Journal of Botany* 102(2), 248–263.

Welker, C., Balasubramanian, V., Petti, C., Rai, K., DeBolt, S., *et al*. (2015b) Engineering plant biomass lignin content and composition for biofuels and bioproducts. *Energies* 8(8), 7654–7676.

Willyard, C. (2018) Expanded human gene tally reignites debate. *Nature* 558, 354–355.

Xie, M., Zhang, J., Tschaplinski, T.J., Tuskan, G.A., Chen, J.G., *et al*. (2018) Regulation of lignin biosynthesis and its role in growth-defense tradeoffs. *Frontiers in Plant Science* 9, 1427.

Xie, S., Ragauskas, A.J. and Yuan, J.S. (2016) Lignin conversion: Opportunities and challenges for the integrated biorefinery. *Industrial Biotechnology* 12(3), 161–167.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., *et al*. (2014) SOAPdenovo-trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12), 1660–1666.

Zeng, Z., Shi, H., Wu, Y. and Hong, Z. (2015) Survey of natural language processing techniques in bioinformatics. *Computational and Mathematical Methods in Medicine* 2015, 674296. DOI: 10.1155/2015/674296

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., *et al*. (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology* 30(6), 549.

Zhang, T., Huang, L., Wang, Y., Wang, W., Zhao, X., *et al*. (2017) Differential transcriptome profiling of chilling stress response between shoots and rhizomes of *Oryza longistaminata* using RNA sequencing. *PloS ONE* 12(11), e0188625.

Zhao, Q. and Dixon, R.A. (2011) Transcriptional networks for lignin biosynthesis: More complex than we thought? *Trends in Plant Science* 16(4), 227–233.

Zhao, Q., Wang, H., Yin, Y., Xu, Y., Chen, F., *et al*. (2010) Syringyl lignin biosynthesis is directly regulated by a secondary cell wall master switch. *Proceedings of the National Academy of Sciences of the USA* 107(32), 14496–14501.

Zhong, R., Richardson, E.A. and Ye, Z.H. (2007) Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. *Planta* 225(6), 1603–1611.

Zhou, J., Lee, C., Zhong, R. and Ye, Z.H. (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *The Plant Cell* 21(1), 248–266.

8 Advances in QTL Mapping and Cloning

Dharminder Bhatia* and Darshan S. Brar
Punjab Agricultural University, Ludhiana, India

Many traits of economic importance in agriculture are quantitative in nature, showing continuous variation. Such traits are controlled by a large number of genes, are highly influenced by environmental factors and often exhibit genotype-by-environment interaction. On account of their large number and relatively small individual effect, these genes were termed as polygenes by Mather (1949) and QTL (quantitative trait loci) by Geldermann (1975). The conventional approaches to the determination of genetic control of such characters include the estimation of total genetic variation, types of gene action and number of genes influencing a particular character.

In the pre-molecular era, quantitative traits were mainly studied through analyses and interpretation of data in the form of means, variance and covariance of relatives. Such an approach provided a conceptual base to partition the total phenotypic variance into genetic and environmental variances and to further divide the genetic variance into additive, dominance and epistatic components of variance. From this information, it became feasible to decide an appropriate breeding strategy and to predict the response of the traits to different methods of selection. It was also possible to estimate the minimum number of genes in the form of 'Effective Factors' controlling a quantitative character. No doubt, this approach has frequently been used satisfactorily in some of the applied plant breeding programmes, but it still falls short of determining the genetic control of quantitative characters with great precision, especially in relation to the exact location and nature of the genes influencing these characters.

The discovery of molecular markers opened up new avenues for determining not only the number of genes more precisely but also the location of the genes in the form of distinctly marked segments of chromosomes designated as QTL. This chapter highlights various advancements relative to mapping and cloning of QTL to better understand genetic architecture of complex quantitative traits and their utilization in genetics and breeding research.

Contemporary History of Quantitative Genetics

The study of quantitative traits found its beginning during the years 1865–1900, following the experiments of Sir Francis Galton, who attempted to elucidate the hereditary principles and published his results in his book *Natural Inheritance*

* Email: d.bhatia@pau.edu

© CAB International, 2020.

in 1889. Galton found that Mendelian principles could not explain the genetic basis of continuous variation and speculated the involvement of blending theory of inheritance for quantitative characters. Despite the perceived failure of Mendelian principles of heredity, Galton provided a strong basis for studying quantitative variation. His work was later expanded by Karl Pearson and his associates, which did not find favour with William Bateson, who was a strong supporter of Mendelian principles.

Thereafter, notable contributions by Johannsen (1909), Nilsson-Ehle (1909) and East (1910) provided evidence of involvement of more than one Mendelian factor and environment in the genetic control and continuous nature of variation exhibited by quantitative characters. Fisher (1918) provided statistical approaches for investigating the average effect of genes in the form of additive, dominance and epistatic components. In addition, the subsequent contributions of Sewell Wright, J.L. Lush, J.B.S. Haldane and several others put quantitative genetics on a firm footing.

The biometrical approaches offered the advantage of analysing the total of variation however limited in distinguishing the effect of individual genes. By 1920, recognition of chromosomal theory of inheritance and phenomenon of linkage and crossing over opened up new possibilities for the study of quantitative variation. The first report of linkage of a quantitative trait (seed size) with a morphological marker (seed coat pigmentation) was reported in common bean (Sax, 1923). Thoday (1961) realized the importance of experimental outcomes of Sax (1923) and encouraged quantitative geneticists to map polygenes using single-gene markers. However, morphological and biochemical markers, such as isozymes, were considered inadequate for scanning the entire genomes for possible location of all genes involved in the control of a quantitative character, which delimited the investigations on quantitative traits.

Molecular Markers: A New Dimension in QTL Mapping

The discovery of abundantly available DNA-based markers and dense molecular maps laid the foundation of QTL analysis that is currently being used. DNA-based markers, being phenotypically neutral in nature, can precisely estimate the effect of linked polygenes (Tanksley, 1993). It was Geldermann who, in 1975, coined the term 'QTL' for a segment of DNA or a region of the genome that has an effect on a quantitative trait.

The limitation posed by the paucity of markers was resolved with the development of restriction fragment-length polymorphisms (RFLPs) as first molecular markers (Botstein, 1980). Thereafter, the landmark publication on QTL detection using the interval mapping approach provided the framework for subsequent QTL mapping (Lander and Botstein 1989) using an RFLP linkage map and log of odds (LOD) score to identify the location and phenotypic effect of QTL. Along with RFLPs, innovation of DNA-sequencing technologies (Sanger and Coulson, 1975), PCR (Mullis *et al.*, 1986) and the subsequent development of molecular markers, such as simple sequence repeats (SSRs) (Litt and Luty, 1989), helped quantitative geneticists to scan the whole of the genome, allowing fine mapping of QTL. These discoveries led to cloning of the first gene underlying a QTL governing fruit size in tomato (Frary *et al.*, 2000), thus validating the relationship of QTL with Mendelian factors. These discoveries ushered in a wave of studies on QTL mapping in plants and animals for a wide range of phenotypic traits. The development of various statistical packages, such as MapMaker/QTL (Lander *et al.*, 1987), Qgene (Nelson, 1997), PLABQTL (Utz and Melchinger, 1996), QTL Cartographer (Wang *et al.*, 2012a), QTLNetwork (Yang *et al.*, 2008) and R/QTL (Broman *et al.*, 2003), accelerated the identification of QTL and their probable phenotypic contribution to the trait variation. These advances helped unfold the genetic architecture of quantitative traits.

Principles of QTL Mapping

The principles of QTL mapping have been discussed in detail by Collard *et al.* (2005). QTL mapping involves locating Mendelian factors in the genome on the basis of linkage disequilibrium between alleles at a marker locus and alleles at linked QTL. The basic principle of QTL

mapping is to classify individuals in mapping populations into marker genotypic classes based on presence and absence of a particular marker locus and to determine whether significant differences exist between classes with respect to the quantitative trait under investigation (see Box 8.1). The QTL mapping involves the development of a mapping population, genotyping that population with a dense collection of molecular markers distributed throughout the genome, phenotyping the population in multi-environments and application of statistical procedures to identify QTL.

Several types of mapping populations, including F_2/F_3, recombinant inbred lines (RILs), doubled-haploids (DHs), backcrosses (BCs), backcross inbred lines (BILs), chromosomal segment substitution lines (CSSLs), multi-parent advanced generation inter-cross (MAGIC) and association mapping panels, are used to map QTL. Any mapping population that is immortal and can be evaluated across different environments will

Box 8.1.

An example of a recombinant inbred line (RIL) population of rice with ten individuals has been taken to demonstrate the QTL analysis using the *t*-test. The ten individuals, along with parents of the RIL population, were genotyped with two SSR markers, M1 and M2, and data were scored as A and B (bands corresponding to each parent). Thousand-grain weight (TGW) of the ten individuals was recorded and QTL analysis performed to determine the association of each of the markers with TGW (a test of single-marker analysis or SMA).

↓Individual/Marker→	M1	M2	TGW (g)
1	B	B	21
2	A	A	32
3	A	A	28
4	B	B	20
5	A	B	29
6	B	A	22
7	A	A	30
8	A	A	29
9	B	B	20
10	B	B	21

For M1 marker:

Step I: Divide individuals into marker classes and find phenotypic mean of each class (A and B).

Individuals 2, 3, 5, 7 and 8 carry band of parent one (A); thus, they belong to Group I, with a mean (m_A) TGW of 29.6 g. Similarly, individuals 1, 4, 6, 9 and 10 carry band of the second parent (B); thus, they belong to Group II, with a mean (m_B) TGW of 20.8 g.

Step II: Find the variance (σ^2_A, σ^2_B) of each group and compute pooled sample variance (σ^2_P).

$$\sigma^2_A = 0.7,\ \sigma^2_B = 2.3$$

$$\sigma^2_P = \frac{(n_A - 1)\sigma^2_A + (n_B - 1)\sigma^2_B}{n_A + n_B - 2} = 1.5$$

Step III: Perform *t*-test and compare observed value with table value at $t_{(0.025,\ nA + nB - 2)}$ for the two-sample *t*-test as follows:

$$t = \frac{m_A - m_B}{\sqrt{\sigma^2_P\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = 11.36$$

The table value of *t* is 2.30, which is smaller than the observed value of *t* (11.36). Therefore, marker M1 is significantly associated with TGW.

be suitable for QTL mapping, though every population has its own merits and demerits (Jaquemin *et al.*, 2013). The size of the mapping population determines the QTL resolution and the number of QTL that can be detected. In general, relatively small-sized populations are unable to detect small-effect QTL. In a study to find QTL for seed oil content in the longest running experiment in maize, a larger number of QTL were identified than those found in previous studies, which was attributable to the larger size of population used (Laurie *et al.*, 2004).

Once the QTL positions in the genome are identified, these can be fine-mapped using QTL-near-isogenic lines (NILs). To develop QTL-NILs, a plant possessing the target phenotype, with minimum size of donor segment in the QTL region and minimum number of residual donor segments throughout the genome, is selected and backcrossed three or four times with the recurrent parent, followed by continuous selfing. Fine-mapping of QTL can be pursued during the development of QTL-NILs at $BC_2F_{2:3}$ or $BC_3F_{2:3}$ stage.

Methods of QTL Analysis

Several methods have been developed for mapping QTL in structured populations, which can be broadly classified as single-marker analysis (SMA) and interval mapping (IM). In SMA, each marker is examined independently for its association with the phenotype, such that '*n*' single marker tests are performed for '*n*' number of markers. SMA is a quick method to detect QTL linked to a marker without estimation of its position and effect. Because SMA does not use the information from linkage maps, the distance between marker and QTL is not known, which may lead to underestimation of the magnitude of the effect of QTL (see Box 8.2). In IM, each pair of adjacent markers is examined separately, such that '$n - 1$' tests are carried out with '*n*' markers. The adjacent markers are taken from the linkage map in order and are linked with each other, which accounts for distance between marker and QTL (Lander and Botstein, 1989). The most widely used IM method is composite interval mapping (CIM). The method is a combination of interval mapping and multiple regression that uses marker interval from IM, along with a subset of well-chosen markers, to identify QTL (Zeng, 1993, 1994). The method reduces the background noise to precisely estimate the effect of QTL, especially when QTL are linked. The success of all of these methods in accurately estimating the location and effect of QTL, depends upon development of precise mapping populations, high-throughput genotyping for developing dense molecular linkage maps and high-precision phenotyping procedures.

Advances in High-throughput Genotyping

RFLPs were the first DNA markers to be successfully used in plants (Helentjaris *et al.*, 1985). RFLPs had many inherent problems, e.g. labour-intensive, time-consuming and risk of exposure to radioactive probes; they were eventually replaced by PCR-based markers, which did not have the problems that RFLPs had. Among them, SSR-based markers were particularly useful. These were relatively inexpensive, abundant in plant genomes and informative. Conventional QTL analysis using PCR-based markers, however, remains time-consuming and labour-intensive, mainly because it requires development of polymorphic markers for linkage analysis.

The improvement of Sanger sequencing throughput the 1990s, in combination with the start of genome sequencing and expressed sequence tag (EST) sequencing programs in model plant species, accelerated the identification of variation at the single-base pair resolution (Wang *et al.*, 1998). Single nucleotide polymorphisms (SNPs) resulting from single-base pair variations are the most abundant DNA markers that are evenly distributed across a genome and these can tag almost any gene or locus in the genome (Brookes, 1999). After 2005, several next-generation sequencing (NGS) platforms, such as Roche 454, Illumina HiSeq2500, ABI 5500xl SOLiD, Ion Torrent, PacBio RS and Oxford Nanopore, emerged. These rapid and cost-effective NGS-based platforms, along with several advances in bioinformatics tools, simplified the discovery of genome-wide SNPs and INDELs (insertions/deletions) (Chen *et al.*, 2013).

Reduced representation techniques, such as RRL (reduced representation libraries),

Box 8.2.

Let us suppose that Q and q are the QTL alleles and M and m are the marker alleles in two parents, respectively. If r is the frequency of recombinant gametes, then non-recombinant gametes will have frequency of 1 – r. Following will be genotypic frequencies and mean of marker and QTL genotype in an RIL population.

	Genotypic frequencies		
Marker genotype	QQ	qq	Mean
MM	1 – r	r	μ_{MM}
mm	r	1 – r	μ_{mm}
Mean	μ_{QQ}	μ_{qq}	

Difference in marker genotype group means will be:

$$\mu_{MM} - \mu_{mm} = [(1-r)\mu_{QQ} + r\mu_{qq}] - [r\mu_{QQ} - (1-r)\mu_{qq}]$$

Solving the above equation, we have:

$$\mu_{MM} - \mu_{mm} = (1-r)(\mu_{QQ} - \mu_{qq})$$

The phenotypic effect is confounded by recombination frequency r in 1 − r. If the marker is tightly linked to QTL, then r = 0, the difference in the phenotype (marker genotype) will be same as the difference at the QTL genotype. However, any value of r will underestimate the effect of the QTL genotype.

CRoPS™ (complexity reduction of polymorphic sequences), RAD-seq (restriction-associated DNA sequencing), GBS (genotyping by sequencing), ddRAD-seq (double-digest restriction-associated sequencing), Rest-seq (restriction DNA sequencing), SLAF-seq (specific-locus amplified fragment sequencing) and Skim GBS (skim-based genotyping by sequencing) harness the power of NGS in identifying tens or hundreds of thousands of markers spread evenly throughout the genome in a large number of individuals (Bhatia *et al.*, 2013; Sun *et al.*, 2013; Golicz *et al.*, 2015). This family of reduced representation techniques is generally referred to as genotype-by-sequencing (GBS) or restriction-site-associated DNA sequencing (RAD-seq). Genotyping with PCR-based markers, such as SSRs, takes weeks to months to complete and involves the cost of marker design as well as assaying, whereas GBS can complete the whole process of genotyping in 2–3 weeks, irrespective of population size (Spindel *et al.*, 2013; Bhatia *et al.*, 2018). The flexible, rapid and low-cost GBS is the ideal choice for genotyping among many researchers for mapping QTL governing complex traits.

Advances in High-throughput Phenotyping

Accuracy of mapping QTL for complex quantitative traits depends entirely upon accurate phenotyping, with little or no effect of environmental conditions. Advancements in phenotyping and genotyping have not been equal; therefore, the progress in genetically dissecting the complex quantitative traits has been slow. Thousands of QTL have been identified in the past for several traits; however, only a handful of QTL have been utilized in mainstream breeding programmes. One of the reasons could be the lack of precision in phenotyping target traits, leading to identification of false positive QTL. False positive QTL represent Type I errors.

In the past decade, efforts have been made to develop new high-throughput platforms/tools for precise and rapid phenotyping of traits on a large scale. These platforms/tools can be used both in controlled and natural environmental conditions, facilitating rapid, dynamic and precise data points. Non-invasive image analysis has enabled rapid phenotyping, not only in the

model plant Arabidopsis but also in several crop plants (Furbank *et al.*, 2011; www.plantphenomics.org.au). The image analysis procedures use different imaging technologies, such as visible imaging (machine vision), imaging spectroscopy (multispectral and hyperspectral remote sensing), thermal infrared imaging, fluorescence imaging, 3D imaging and tomographic imaging, such as magnetic resonance imaging (MRI), positron emission detectors for short-lived isotopes (PET) and X-ray computed tomography (CT), for collecting data on complex quantitative traits related to plant growth, yield and adaptation to biotic or/and abiotic stresses (Li *et al.*, 2014). Based on these technologies, several controlled-environment and field-based phenotyping platforms have been developed that combine the advancements in sensing technologies, aeronautics and computing. Much of the focus has also been on improving field-level phenotyping procedures to enhance precision and to control field-based environmental variation. Such procedures are now being increasingly recognized as high-throughput tools for accurately phenotyping large numbers of plant populations (White *et al.*, 2012).

Advances in Mapping QTL

Recent advances in mapping and cloning of QTL are utilizing the power of NGS platforms and development of improved statistical algorithms. Advanced genotyping platforms, such as GBS and whole-genome resequencing, have accelerated the identification of QTL, with greater precision and power (Bhatia *et al.*, 2018). QTL mapping requires scanning the whole genome with a dense set of molecular markers. The process becomes time-consuming, laborious and economically inefficient with large population sizes. Bulked segregants analysis (Michelmore *et al.*, 1991) and selective genotyping (Darvasi and Soller, 1994; Sun *et al.*, 2010) represent both time- and cost-saving strategies that have simplified the analytical process of QTL mapping. Quantitative traits show normal distribution with two extreme tails. All the individuals belonging to each tail can be analysed as a pool in bulked segregants analysis (BSA), and individually in selective genotyping.

Modified BSA strategies for mapping QTL

Several BSA-based modifications have been developed to rapidly locate the target genes/QTL. These modifications use increased tail size from a large population and high-density markers, obviating the need to validate the putative markers by genotyping entire populations (Zhou *et al.*, 2016; Nguyen *et al.*, 2019). As a consequence, there is a dramatic reduction in genotyping cost, whereas the statistical power in QTL mapping is comparable to the entire population analysis (Macgregor *et al.*, 2008; Sun *et al.*, 2010; Vikram *et al.*, 2012). Besides the availability of whole-genome sequences of several important crop plants, developments in the field of bioinformatics have facilitated the functional validation and cloning of QTL. This has led to cloning of several QTL in the past decade (Table 8.1), enabling researchers to understand the genetic architecture of several economically important traits. Advances in QTL-mapping strategies can be categorized into: (i) those that combine the power of NGS and BSA; and (ii) those that exploit historical recombination events in natural populations.

QTL-Seq: an extension of BSA

BSA has been a widely used technique to rapidly identify the linkage of molecular markers with the causal phenotype. First described by Michelmore *et al.* (1991), the technique is based on the principle that the bulk of DNA sharing a similar phenotype must contain the same allele; therefore, two different bulks may be able to identify the linkage of allele polymorphism with the difference in phenotype. QTL-seq is an extension of BSA that uses the power of NGS. It has been introduced for rapid identification of plant QTL by whole-genome resequencing of DNA bulks, each composed of 10–20 individuals showing extreme opposite trait value, from a segregating progeny (Takagi *et al.*, 2013a).

For QTL mapping using QTL-seq, first a mapping population is generated by crossing two genotypes showing contrasting phenotypes for the trait(s) of interest. The mapping population is phenotyped for targeted traits, preferably in replications and in multiple environments. DNA is extracted from 10 to 20 individuals

Table 8.1. Some examples of cloned quantitative trait loci (QTL) for yield-related traits using positional cloning in rice.

Trait	Name of QTL	NCBI accession number/ RGAP Locus id	Reference(s)
Grain number	*GN1a*	LOC_Os01g10110.1	Ashikari *et al*. (2005)
	Ghd7	EU286800, EU286801	Xue *et al*. (2008)
	DEP1	FJ039904, FJ039905	Huang *et al*. (2009)
	DEP3	LOC_Os06g46350.1	Qiao *et al*. (2011)
	Ghd8/DTH8	LOC_Os08g07740.1	Wei *et al*. (2010); Yan *et al*. (2011)
	NOG1	MF687920, MF687921	Huo *et al*. (2017)
Grain number/ culm strength	*SCM2/APO1*	LOC_Os06g45460.1	Ikeda *et al*. (2007); Ookawa *et al*. (2010); Terao *et al*. (2010)
Panicle size	*SP1*	LOC_Os11g12740.1	Li *et al*. (2009)
Grain filling	*GIF1*	U87973	Wang *et al*. (2008)
Tiller number and grain number	*EP3*	LOC_Os02g15950.1	Piao *et al*. (2009)
	IPA1/WFP/OsSPL14	LOC_Os08g39890.1	Jiao *et al*. (2010); Miura *et al*. (2010)
Grain length	*GS3*	DQ355996	Fan *et al*. (2006)
Grain width	*GW2*	EF447275	Song *et al*. (2007)
	qSW5/GW5	AB433345	Shomura *et al*. (2008)
	GS5	LOC_Os05g06660.1, LOC_Os05g06660.2	Li *et al*. (2011)
	GW8	JX867117	Wang *et al*. (2012b)
	TGW6		Ishimaru *et al*. (2013)
Plant architecture	*PROG1*	FJ155665	Tan *et al*. (2008)
Root architecture	*Dro1*	AB689741, AB689742	Uga *et al*. (2013)
Panicle architecture	*OsLG1*	AB776991	Ishii *et al*. (2013)

showing extreme phenotypic values and bulked to generate a highest bulk and a lowest bulk. Each bulk is sequenced with >6× genome coverage on a suitable NGS platform. NGS short reads of bulks are aligned to the reference genome and SNPs are called. Both the bulks should have genomes from both the parents in a 1 : 1 ratio at most of the genomic regions; however, it may differ at the position of QTL in the genome. This can be identified using SNP-index and ΔSNP-index. The SNP-index is the proportion of reads harbouring the SNPs that are different from the reference genome sequence. ΔSNP-index is obtained by subtraction of the SNP-index of the highest bulk from that of the lowest bulk (Takagi *et al*., 2013a). The SNP-index = 1, if all the short reads at a genomic position contain SNPs from parent 1. The SNP-index = 0, if all the short reads represent SNPs from the other parent. The SNP index of both the bulks with a particular parent is calculated and plotted in the form of ΔSNP index using a sliding window approach. Contrasting patterns of SNP index (1–0) in both the bulks will reveal the putative genomic region harbouring the QTL. Till now, this approach has been used to identify QTL for important traits in several crop plants, including rice (Takagi *et al*., 2013a; Kadambari *et al*., 2018), tomato (Ruangrak *et al*., 2018), *Brassica* (Shu *et al*., 2018) and soybean (Zhang *et al*., 2018).

MutMap

MutMap is a rapid method of identification of major gene/QTL in the F_2 population derived from a cross between a mutant and wild type (Abe *et al*., 2012). The method makes available the mutant plant and associated marker to the plant breeder to accelerate a crop improvement programme.

The MutMap method requires a homozygous recessive mutant phenotype that needs to be crossed to its wild type to generate an F_2 population. First a mutagen, such as ethyl methane sulphonate (EMS), is used to mutagenize a selected line of crop plant that has a reference

genome sequence. Mutagenized plants (M_1) are self-pollinated to generate M_2. In M_2 or later generations, a mutant for the target trait can be identified and self-pollinated to generate a homozygous mutant phenotype. Once the mutant is identified, it is crossed with the wild-type plant of the same line as used for mutagenesis. The F_1 plant is self-pollinated to generate F_2 progeny. The number of segregating loci governing a targeted phenotype in the F_2 will be minimal as the F_2 progeny are derived from a cross between the mutant and its parental wild-type plant. In most cases, the targeted mutant phenotype may be attributable to a single-gene mutation, which can be clearly observed even if the phenotypic difference is small. Since the mutant phenotype is controlled by few genes, phenotyping only a small population will show the segregation of the required phenotype. The DNA of ≥20 F_2 progeny showing the mutant phenotype is bulked and sequenced using the NGS platform with >10× coverage. The NGS reads are aligned to the reference genome and SNPs are identified. The SNP index is calculated for each SNP. The SNP index will be 1 or near 1 for causal SNP, whereas it will be 0.5 for any unlinked loci. The locus with SNP index will be the causal genomic region governing the phenotype.

MutMap has been used to identify the causal mutation responsible for pale-green leaves and semi-dwarfness in rice (Abe *et al.*, 2012). This method has been used in a study by Takagi *et al.* (2015), where a loss of function mutation in *OsRR22* was identified that was responsible for salt tolerance in Hitomebore salt tolerant 1 (*hst1*), an EMS-based mutant of cv. Hitomebore. The mutant was later used to release the improved variety 'Kaijin' in just 2 years. The Kaijin differed from the wild-type Hitomebore by 201 SNPs, with similar yield and growth level, restoring rice production in the tsunami-affected areas of Japan (Takagi *et al.*, 2015).

MutMap-Gap

This approach is a variation of the MutMap scheme, which involves identification of a candidate region harbouring a mutation of interest using the MutMap, followed by *de novo* assembly, alignment and identification of the mutation within genome gaps (Takagi *et al.*, 2013b). In MutMap, the genomic fragment responsible for the mutant phenotype should be present in the reference genome to be tapped for causal mutation. However, it may not always be the case. Similarly, MutMap will have a limitation if the reference genome has gaps at the position of causal mutation. MutMap-Gap combines the MutMap with targeted gap filling by *de novo* assembly of the wild-type genome. Because of the availability of *de novo* genomic assembly of the wild parental line, the technique was able to identify a blast resistance *Pii* gene in rice variety Hitomebore, which was absent in the Nipponbare reference genome (Takagi *et al.*, 2013b).

MutMap-Gap first applies MutMap to identify a candidate genomic interval based on SNP index. This candidate genomic interval may be located in the gap region of the reference genome, so a *de novo* assembly of wild-type genome as scaffold is generated. The NGS reads of F_2 progeny are again aligned to *de novo* assembly of wild type. Next, SNPs are identified and SNP index is calculated for each SNP. The combined SNP index analysis from MutMap and MutMap-Gap will be able to identify the causal mutation and candidate genomic region.

MutMap-based approaches integrate the classical mutant analysis and NGS methods. The availability of major mutation for any complex trait will be the prime requirement for these approaches to be successful.

BSR-seq

BSR-seq is a BSA-based mapping strategy that uses RNA-seq data (Liu *et al.*, 2012). The technique is highly suitable for species with relatively large genomes containing much repetitive DNA, which increases the cost of whole-genome DNA sequencing. The BSR-seq identifies the mapping interval associated with the mutant phenotype. The differential expression of genes within the BSR-seq-identified mapping interval can then be used to identify candidate genes responsible for the mutant phenotype. In addition, the approach results in the detection of expression QTL (eQTL) and polymorphic SNPs tightly linked to the mutant phenotype, which facilitates fine mapping and cloning of gene/QTL. Unlike RNA-seq, BSR-seq uses unreplicated data to map QTL, though replicated analysis will have the added advantage of precisely identifying the differentially expressed genes. The kind of tissue for

RNA-seq analysis in BSR-seq does not have an impact on the mapping result much, but will affect the results of differentially expressed genes. Because even if causal candidate gene expression is absent in the sample tissue used for RNA-seq analysis, the SNPs linked with the causal candidate genes that are expressed in the RNA-seq analysis will be used as markers to map the causal gene. However, for the dual advantage of mapping and identifying differentially expressed candidate genes, it would be ideal to use the tissue in which the candidate gene is expressed for RNA-seq analysis (Liu *et al.*, 2012).

BSR-seq has been successfully used for mapping stipe rust-resistance loci *YrMM58* and *YrHY1* on chromosome 2AS (Wang *et al.*, 2018) and leaf senescence gene *els1* on chromosome 2BS (Li *et al.*, 2018) in a segregating biparental wheat population. In maize, BSR-seq has been used to identify SNPs associated with differentially expressed genes for waterlogging in a pool of ten sensitive and eight tolerant inbred lines (Du *et al.*, 2017).

Genome-wide association: mapping QTL in natural population

Mapping in biparental population or linkage mapping has limitations because of: (i) time and cost required to develop a mapping population; (ii) limited recombination events; (iii) segregation of only alleles present in the parents; (iv) crossability issues in a few wide crosses and some vegetatively propagated crops, which hinders the development of mapping populations; and (v) time and skill required to develop mapping populations in forest and fruit crops. In addition, biparental populations can identify a limited set of QTL for a particular trait. The genetic architecture of quantitative traits suggests that a subset of QTL for a trait segregates in the natural population. Genome-wide association studies (GWAS), which have the ability to dissect variations in the natural population, is a powerful approach to identify the QTL subsets. In general, linkage mapping identifies QTL in a window of 10–20 cM because of limited recombination events. GWAS instead takes advantage of historical recombination events in natural populations (a collection of more diverse lines) to resolve the QTL down to the sequence level (Zhu *et al.*, 2008). Typically, the goal of GWAS is to understand variation attributable to complex traits by finding association between the genotype of a large number of markers (SNPs) and the phenotype. Following its demonstration in crop plants (Thornsberry *et al.*, 2001) and with a sharp decline in the cost of sequencing, GWAS has accentuated interest for its use among the scientific community. The approach has been widely utilized in several crop plants to map QTL for several important traits, such as yield component traits in rice and wheat (Huang *et al.*, 2010; Neumann *et al.*, 2011), fusarium head blight resistance in barley (Massman *et al.*, 2011), and oleic acid content and resistance to southern leaf blight in maize (Belo *et al.*, 2008; Kump *et al.*, 2011).

GWAS relies on linkage disequilibrium (LD) to identify the relationship between phenotype and genotype (Flint Gracia *et al.*, 2003) and well connects the advances in genotyping technologies. The mapping population in GWAS can be a collection of wild species, landraces, cultivars from multiple programmes, geographically distinct regions or regional breeding programmes. However, collection of intra-species diversity is recommended to generate a well-sampled collection known as an association panel, which defines the robustness of an association panel to identify the genetic basis of any variation. Recently, the development of well-sorted association panels, such as the 2K association panel (McCouch *et al.*, 2016) and the 3K association panel (The 3,000 rice genome project, 2014) in rice are helping to tap immense genetic variation down to precise QTL levels.

LD is the non-random association of alleles at different loci in the population and varies in self-pollinated as well as in cross-pollinated crops. LD is generated in the populations by several factors that affect Hardy–Weinberg equilibrium in plants. The extent of LD that exists in a population around a particular locus determines the resolution of mapping QTL. In addition, the size of the population, population structure, relatedness in the population and minor allele frequency (MAF) are the crucial factors in mapping QTL. Advanced genotyping technologies that generate genome wide marker (basically SNPs and InDels) information are the optimum choice to capture LD between marker and the QTL.

Consensus QTL detection using meta QTL analysis

QTL mapping studies in biparental populations reveal only a portion of the genetic architecture of a trait since these mapping populations segregate for a very few QTL of large effects and a very limited number of QTL with small effects (Salvi and Tuberosa, 2005; Holland, 2007). In addition, very few QTL have been used in marker-assisted selection (MAS), because of lack of precision in locating QTL that explain only a small proportion of variation probably because of QTL × environment interaction. For MAS, one major requirement is the smallest QTL region to avoid possible undesirable effects of genes carried by an introgressed genomic segment. Further, there is little congruency in detection of QTL in different populations in different studies. This may be attributable to different subsets of QTL segregating in different populations (Holland, 2007) and differences in family structure, sample size, marker maps or QTL detection method used. However, data from multiple populations, representing different sampling variation, can be taken into account to narrow down the QTL region.

Meta-analysis of QTL is one approach that can pool data from different studies on the same or a similar trait to build a consensus map of the QTL region, thus providing an opportunity to narrow down the QTL region and seek candidate genes. There have been a number of QTL mapping studies in the past decade. Data from most of these studies can be assessed from public QTL databases, such as Gramene (www.gramene.org). Meta QTL analysis using data from one's own experiment and from public databases will increase the statistical power of QTL detection and precision of effect estimation. Different meta-analysis studies have been proposed to study the congruency of QTL from different studies. These include a meta-analysis study based on model selection (Goffinet and Gerber, 2000), a meta-analytic approach to overcome the between-studies heterogeneity and to refine both QTL location and the magnitude of the genetic effects (Etzel and Guerra, 2002), and a two-stage meta-analysis to integrate multiple independent QTL-mapping experiments, implemented using the MetaQTL package (Veyrieras *et al.*, 2007). These methods have resulted in several QTL meta-analysis studies in the past 10–20 years (Ioannidis *et al.*, 2001, Rong *et al.*, 2007, Truntzler *et al.*, 2010).

Positional cloning of QTL

QTL cloning begins with identification of individual QTL (Mendelian component) in the genome using QTL-mapping procedures. This is followed by fine-mapping and cloning of each QTL. Since QTL for a target trait will be more than one, most of the studies target to clone a major-effect QTL explaining >20% of phenotypic variation. For fine-mapping a QTL, a plant possessing the target phenotype, with minimum size of donor segment at QTL region and minimum number of residual donor segments throughout the genome, is selected and backcrossed three or four times with the recurrent parent, followed by continuous selfing to develop QTL-NILs. During each backcross, QTL is delimited to a narrower region by genotyping the population by developing more markers from the QTL region, phenotyping and conducting QTL analysis or recombinant analysis. Separate QTL-NILs need to be developed for cloning individual QTL referred to as Mendelizing QTL. Individual QTL-NIL will then be crossed to the recurrent parent to generate a large F_2 population (>2000) to increase the mapping resolution (Salvi and Tuberosa, 2007). In that population, the difference in the QTL region is responsible for any phenotypic differences attributable to the absence of any other genetic differences. In addition, a large mapping population will provide recombinants to confine the QTL region to a single-gene region.

Positional cloning of QTL involves analysis of a large number of recombinants, enriching the recombinant region with molecular markers by chromosome walking and generating a physical map of the recombinant region covering the gene of interest. The candidate gene in the QTL region is cloned, followed by complementation by transformation and *de novo* determination of the gene sequence. The first QTL cloned by positional cloning was the fruit size QTL '*fw2.2*' in tomato (Frary *et al.*, 2000), which used YAC (yeast artificial chromosome), cDNA and cosmid libraries, followed by genetic complementation analysis of transgenic plants. With the availability of the whole genome sequence of crop plants, the amount of effort in fine-mapping and

positional cloning of QTL has dramatically decreased. The whole genome sequence has made possible conversion of a genetic region of initial mapping to a physical region in the genome and selection of putative candidate genes in the genome, eliminating the laborious task of chromosome walking. The identification of putative candidate genes helps to design the markers from the target region only and analysing the population for harbouring the narrower region of the QTL. This exponentially increases the cloning of QTL conditioning important traits. With the availability of the whole genome sequence, more than 100 QTL have been cloned in rice using positional cloning after 2005. Similar trends can be seen in Arabidopsis, maize and other crops such as wheat in the coming years. Furthermore, advances in high-throughput genotyping platforms, development of several BSA-NGS-based strategies and exploiting the natural variation through GWAS will boost the positional cloning of QTL of economically important traits and understanding the genetic architecture of complex quantitative traits.

Homology-based cloning of QTL

Though positional cloning has accelerated with advances in genomics and bioinformatics, it remains an arduous task to clone a QTL. However, the genomics era has provided ample opportunities to clone orthologous QTL using the information on cloned QTL from related species by comparative genomic approaches. Homology-based cloning is a powerful approach for the discovery of gene/QTL on the basis of collinearity of genes. Recently, taking advantage of cloned yield-related QTL in rice, this approach has been successfully used in cloning several yield-related QTL in wheat, e.g. *TaGW2* (Su *et al.*, 2011), *TaSus2* (Jiang *et al.*, 2011), *TaCwi-A1* (Ma *et al.*, 2012), *TaGS1a* (Guo *et al.*, 2013), *TaGS-D1* (Zhang *et al.*, 2014), *TaTGW6-A1* (Hanif *et al.*, 2016) and *TaFlo2-A1* (Sajjad *et al.*, 2017). The cloning of these QTL has led to the development of functional markers that are being used in MAS. The first step in this approach is to identify orthologous sequences of cloned QTL from the related species in the target genome using the BLAST analysis. The orthologous gene sequences are used to design gene-specific, overlapping primers to amplify full-length genes in different lines of target species with contrasting traits. Further, gene sequences amplified in different lines are aligned to identify nucleotide differences. The putative nucleotide differences are used to develop functional markers, which are then validated in diverse germplasm lines showing variation for the target trait. The application of this approach can be extended to distantly related species. As an example, wheat *Rht1* gene and maize dwarfing gene *D8* were isolated using the *Arabidopsis GA1* gene (Peng *et al.*, 1999).

Homology-based cloning approaches have an added advantage of boosting positional cloning of QTL by enriching the molecular maps of target species by getting markers from orthologous regions from related species. In wheat, candidate genes of the *Ph1* (*pairing homoeologous 1*) locus has been identified by using this approach. Comparative mapping of wheat 5BL-specific EST (*Ph1* region) onto the rice chromosome identified the orthologous region on rice chromosome 9. Careful analysis of rice genes located on the rice orthologous region identified nine candidate genes that could be the target for cloning the *Ph1* locus in wheat (Sidhu *et al.*, 2008). Further, silencing of one of the genes (*metaphase I- specific* gene) in wheat resulted in a similar phenotype to that of the *Ph1* gene, thus identifying the putative candidate of the *Ph1* gene in wheat (Bhullar *et al.*, 2014).

Future Research Priorities in QTL Mapping and Cloning

1. So far, a major emphasis has been placed on mapping QTL with larger effects. However, there is an urgent need to look for minor-effect QTL. With the availability of a vast array of genomic tools, it is now possible to capture minor-effect QTL, followed by their pyramiding to enhance the trait effect. As an example, emphasis may be given to identification of minor QTL for tolerance to stem borer in rice, cotton and pulses, sheath blight in rice, salinity and drought tolerance at reproductive stage and pyramiding them in a background for which major QTL could not be identified.

2. Emphasis should be given to development of specialized genetic stocks, such as CSSLs and NILs, for mapping and cloning of QTL. These stocks can be crossed with the recurrent parent to develop larger populations for fine-mapping and cloning of QTL. These may further help to understand QTL × QTL interactions.

3. Most of the QTL-mapping studies have focused on identification of QTL in a single genetic background. However, the expression of QTL changes when transferred into a different genetic background. The future priorities should be to identify robust and reproducible QTL that are stable across different genetic backgrounds.

4. Many institutes in developing countries have inadequate facilities for precise phenotyping, especially complex quantitative traits, such as tolerance to drought and heat. Additional funding is requited to develop high-throughput phenotyping facilities, particularly in the context of changing climatic conditions that adversely affect crop productivity and sustainability.

5. There are several successful examples of pyramiding major genes; however, negligible efforts have been made to pyramid QTL controlling a trait. As an example, several QTL of economically important traits, such as QTL for yield component traits and drought, have been mapped and cloned in important crop plants, which should be transferred into elite backgrounds to have enhanced trait effect and to study QTL × QTL interactions.

6. Resistance (R) genes play a key role in host-plant resistance. However, the deployment and breakdown of R-genes have been a frustrating battle for plant breeders and pathologists; thus, durable, field resistance is needed. On the other hand, quantitative resistance loci (QRL) have been suggested to provide durable resistance in crop plants. Future research may be focused on mapping and cloning QRL and evaluating their durability in crop protection.

7. With the availability of the whole genome sequence of most of the important crop plants, comparative genomics promises to accelerate the cloning of QTL governing complex quantitative traits. Future research can be accelerated to map and clone QTL of economic importance by using the information on QTL from related species.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., *et al*. (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnol*ogy 30, 174–178.

Ashikari, M., Sakakibara, H., Lin, S.Y., Yamamoto, T., Takashi, T., *et al*. (2005) Cytokinin oxidase regulates rice grain production. *Science* 309, 741–745.

Belo, A., Zheng, P.Z., Luck, S., Shen, B., Meyer, D.J., *et al*. (2008) Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics* 279, 1–10.

Bhatia, D., Wing, R.A., Yu, Y., Chougule, K., Kudrna, D., *et al*. (2018) Genotyping by sequencing of rice interspecific backcross inbred lines identifies QTLs for grain weight and grain length *Euphytica* 214, 41. DOI: 10.1007/s10681-018-2119-1.

Bhatia, D., Wing, R.A. and Singh, K. (2013) Genotyping by sequencing, its implications and benefits. *Crop improvement* 40, 101–111.

Bhullar, R., Nagarajan, R., Bennypaul, H., Sidhu, G.K., Sidhu, G., *et al*. (2014) Silencing of metaphase I-specific gene results in a phenotype similar to that of the pairing homeologous I (*Ph1*) gene mutations. *Proceedings of the National Academy of Science*s *of the USA* 39, 14187–14192.

Botstein, D., White, R.L., Skolnick, M. and Davism, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *The American Journal of Human Genetics* 32, 314–331.

Broman, K.W., Wu, H., Sen, S. and Churchill, G.A. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890.

Brookes, A.J. (1999) The essence of SNPs. *Gene* 234, 177–186.

Chen, H., He, H., Zhou, F., Yu, H. and Deng, X.W. (2013) Development of genomics-based genotyping platform and their application in rice breeding. *Current Opinion in Plant Biology* 16, 247–254.

Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B. and Pang, E.C.K. (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142, 169–196.

Darvasi, A. and Soller, M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138, 1365–1373.

Du, H., Zhu, J., Su, H., Huang, M., Wang, H., Ding, S., *et al*. (2017) Bulked Segregant RNA-seq reveals differential expression and SNPs of candidate genes associated with waterlogging tolerance in maize. *Frontiers in Plant Science* 8, 1022. DOI: 10.3389/fpls.2017.01022.

East, E.M. (1910) A Mendelian interpretation of variation that is apparently continuous. *The American Naturalist* 44, 65–82.

Etzel, C.J. and Guerra, R. (2002) Meta-analysis of genetic-linkage analysis of quantitative-trait loci. *The American Journal of Human Genetics* 71, 56–65.

Fan, C.C., Xing, Y.Z., Mao, H.L., Lu, T.T., Han, B., *et al*. (2006) *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics* 112, 1164–1171.

Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S. (2003) Structure of linkage disequilibrium in plants. *Annual Reviews of Plant Biology* 54, 357–374.

Frary, A., Nesbitt, T.C., Frary, A., Grandillo, S., Knaap, E., *et al*. (2000) Fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289, 85–88.

Furbank, R.T. and Tester, M. (2011) Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16, 635–644.

Geldermann, H. (1975) Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and applied Genetics* 46, 319–330.

Goffinet, B. and Gerber, S. (2000) Quantitative trait loci: A meta-analysis. *Genetics* 155, 463–473.

Golicz, A.A., Bayer, P.E. and Edward, D. (2015) Skim-based genotyping by sequencing. In: Batley, J. (ed.) *Plant Genotyping: Methods and Protocols*, *Methods in Molecular Biology*, Volume 1245. Springer, New York, pp. 257–270.

Guo, Y., Sun, J., Zhang, G., Wang, Y., Kong, F., *et al*. (2013) Haplotype, molecular marker and phenotype effects associated with mineral nutrient and grain size traits of TaGS1a in wheat. *Field Crop Research* 154, 119–125.

Hanif, M., Gao F., Liu, J., Wen, W., Zhang, Y., *et al*. (2016) TaTGW6-A1, an ortholog of rice TGW6, is associated with grain weight and yield in bread wheat. *Molecular Breeding* 36, 1. DOI: 10.1007/s11032-015-0425-z.

Helentjaris, T., King, G., Slocum, M., Siedenstrang, C. and Wegman, S. (1985) Restriction fragment length polymorphism as probes for plant diversity and their developments as tools for applied plant breeding. *Plant Molecular Biology* 5, 109–118.

Holland, J. (2007) Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* 10, 156–161.

Huang, X.H., Wei, X.H., Sang, T., Zhao, Q.A., Feng, Q., *et al*. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* 42, 961–976.

Huang, X.Z., Qian, Q., Liu, Z.B., Sun, H., He, S.Y., *et al*. (2009) Natural variation at the *DEP1* locus enhances grain yield in rice. *Nature Genetics* 41, 494–497.

Huo, X., Wu, S., Zhu, Z., Liu, F., Fu, Y., *et al*. (2017) NOG1 increases grain production in rice. *Nature Communications* 8, 1497.

Ikeda, K., Ito, M., Nagasawa, N., Kyozuka, J. and Nagato, Y. (2007) Rice *ABERRANT PANICLE ORGANIZATION 1*, encoding an F-box protein, regulates meristem fate. *Plant Journal* 51, 1030–1040.

Ishii, T., Numaguchi, K., Miura, K., Yoshida, K., Thanh, P.T., *et al*. (2013) *OsLG1* regulates a closed panicle architecture in domesticated rice. *Nature Genetics* 45, 462–465.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. and Contopoulos-Ioannidis, D.G. (2001) Replication validity of genetic association studies. *Nature Genetics* 29, 306–309.

Ishimaru, K., Hirotsu, N., Madoka, Y., Murakami, N., Hara, N., *et al*. (2013) Loss of function of the IAA-glucose hydrolase gene TGW6 enhances rice grain weight and increases yield. *Nature Genetics* 45, 707–711.

Jacquemin, J., Bhatia, D., Singh, K. and Wing, R.A. (2013) The International *Oryza* Map Alignment Project: Development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Current Opinion in Plant Biology* 16, 147–156.

Jiang, Q., Hou, J., Hao, C., Wang, L., Ge, H., *et al*. (2011) The wheat (*T. aestivum*) sucrose synthase 2 gene (*TaSus2*) active in endosperm development is associated with yield traits. *Functional and Integrated Genomic* 11, 49–61.

Jiao, Y.Q., Wang, Y.H., Xue, D.W., Wang, J., Yan, M.X., *et al*. (2010) Regulation of *OsSPL14* by *OsmiR156* defines ideal plant architecture in rice. *Nature Genetics* 42, 541–544.

Johannsen, W. (1909) *Elemente der exakten Erblichkeitslehre*. [Elements of an Exact Theory of Heredity.] Gustav Fischer, Jena, Germany.

Kadambari, G., Vemireddy, L.R., Srividhya, A., Nagireddy, R., Jena, S.S., Gandikota, M., *et al*. (2018) QTL-Seq-based genetic analysis identifies a major genomic region governing dwarfness in rice (*Oryza sativa* L.). *Plant Cell Reports* 37(4), 677–687.

Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., *et al*. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* 43, 163–168.

Lander, E.S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185–199.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., *et al*. (1987) MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1 (2), 174–181.

Laurie, C.C., Chasalow, S.D., LeDeaux, J.R., McCarrroll, R., Bush, D., *et al*. (2004) The genetic architecture of response to long-term artificial selection for oil concentration in maize kernel. *Genetics* 168, 2141–2155.

Li, L., Zhang, Q. and Huang, D. (2014) A review of imaging techniques for plant phenotyping. *Sensors* 14, 20078–20111.

Li, M., Li, B., Guo, G., Chen, Y., Xie, J., *et al*. (2018) Mapping a leaf senescence gene els1 by BSR-Seq in common wheat. *The Crop Journal* 6, 236–243.

Li, S.B., Qian, Q., Fu, Z.M., Zeng, D.L., Meng, X.B., *et al*. (2009) *Short panicle 1* encodes a putative PTR family transporter and determines rice panicle size. *The Plant Journal* 58, 592–605.

Li, Y.B., Fan, C.C., Xing, Y.Z., Jiang, Y.H., Luo, L.J., *et al*. (2011) Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nature Genetics* 43, 1266–1269.

Litt, M. and Luty, J.A. (1989) A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *The American Journal of Human Genetics* 44, 397–401.

Liu, S., Yeh, C.-T., Tang, H.M., Nettleton, D. and Schnable, P.S. (2012) Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS ONE* 7, e36406.

Ma, D., Yan, J., He, Z., Wu, L. and Xia, X. (2012) Characterization of a cell wall invertase gene *TaCwi-A1* on common wheat chromosome 2A and development of functional markers. *Molecular Breeding* 29, 43–52.

Macgregor, S., Zhao, Z.Z., Henders, A., Martin, N.G., Montgomery, G.W., *et al*. (2008) Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Research* 36, e35.

Massman, J., Cooper, B., Horsley, R., Neate, S., Dill-Macky, R., *et al*. (2011) Genome-wide association mapping of fusarium head blight resistance in contemporary barley breeding germplasm. *Molecular Breeding* 27, 439–454.

Mather, K. (1949) *Biometrical Genetics*. Methuen, London.

McCouch, S., Wright, M., Tung, C-W., Maron, L., McNally, K., *et al*. (2016) Open access resources for genome wide association mapping in rice. *Nature Communications* 7, 10532.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences of the USA* 88, 9828–9832.

Miura, K., Ikeda, M., Matsubara, A., Song, X.J., Ito, M., *et al*. (2010) *OsSPL14* promotes panicle branching and higher grain productivity in rice. *Nature Genetics* 42, 545–549.

Mullis, K.B., Faloona, F.A., Scharf, S.J., Saiki, R.K., Horn, G.T., *et al*. (1986) Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology* 51, 263–273.

Nelson, J.C. (1997) QGENE: Software for marker-based genomic analysis and breeding. *Molecular Breeding* 3, 239–245.

Neumann, K., Kobiljski, B., Dencic, S., Varshney, R.K. and Borner, A. (2011) Genome-wide association mapping: A case study in bread wheat (*Triticum aestivum* L.). *Molecular Breeding* 27, 37–58.

Nguyen, K.L., Grondin, A., Courtois, B. and Gantet, P. (2019) Next-generation sequencing acceleraytes crop genome discovery. *Trends in Plant Science* 24, 263–274.

Nilsson-Ehle H (1909) *Kreuzunguntersuchungen an Hafer und Weizen*. H. Ohlssons Buchdruckerei, Lund, Sweden.

Ookawa, T., Hobo, T., Yano, M., Murata, K., Ando, T., *et al*. (2010) New approach for rice improvement using a pleiotropic QTL gene for lodging resistance and yield. *Nature Communications* 132, 1–11.

Peng, J.R., Richards, D.E., Hartley, N.M., Murphy, G.P., Devos, *et al*. (1999) 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400, 256–261.

Piao, R.H., Jiang, W.Z., Ham, T.H., Choi, M.S., Qiao, Y.L., *et al*. (2009) Map-based cloning of the *ERECT PANICLE 3* gene in rice. *Theoretical and Applied Genetics* 119, 1497–1506.

Qiao, Y.L., Piao, R.H., Shi, J.X., Lee, S.I., Jiang, W.Z., *et al*. (2011) Fine mapping and candidate gene analysis of *dense and erect panicle 3*, *DEP3*, which confers high grain yield in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* 122, 1439–1449.

Rong, J., Feltus, F.A., Waghmare, V.N., Pierce, G.J., Chee, P.W., *et al*. (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fibre development. *Genetics* 176, 2577–2588.

Ruangrak, E., Su, X., Huang, Z., Wang, X., Guo, Y., *et al*. (2018) Fine mapping of a major QTL controlling early flowering in tomato using QTL-seq. *Canadian Journal of Plant Science* 98 (3), 672–682.

Sajjad, M., Ma, X., Khan, S.H., Shoaib, M., Song, Y., *et al*. (2017) *TaFlo2-A1*, an ortholog of rice *Flo2*, is associated with thousand grain weight in bread wheat (*Triticum aestivum* L.). *BMC Plant Biology* 17, 164. DOI: 10.1186/s12870-017-1114-3

Salvi, S. and Tuberosa, R. (2007) Cloning QTLs in plants. In: Varshney, R.K. and Tuberosa, R. (eds) *Genomics-Assisted Crop Improvement: Volume 1: Genomics Approaches and Platforms*. Springer, New York, pp. 207–225.

Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs: Present and future challenges. *Trends in Plant Science* 10, 297–304.

Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94, 441–448.

Sax, K. (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8, 552–560.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., *et al*. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nature Genetics* 40, 1023–1028.

Shu, J., Liu, Y., Zhang, L., Li, Z., Fang, Z., *et al*. (2018) QTL seq for rapid identification of candidate genes for flowering time in broccoli × cabbage. *Theoretical and Applied Genetics* 131 (4), 917–928.

Sidhu, G.K., Rustgi, S., Shafqat, M.N., Wettstein D.V. and Gill K.S. (2008) Fine structure mapping of a gene rich region of wheat carrying Ph1, a suppressor of crossing over between homeologous chromosomes. *Proceedings of the National Academy of Sciences of the USA* 15, 5815–5820.

Song, X.J., Huang, W., Shi, M., Zhu, M.Z. and Lin, H.X. (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genetics* 39, 623–630.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., *et al*. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations *Theoretical and Applied Genetics* 126, 2699–2716.

Su, Z., Hao, C., Wang, L., Dong, Y. and Zhang, X. (2011) Identification and development of a functional marker of *TaGW2* associated with grain weight in bread wheat (*Triticum aestivum* L.) *Theoretical and Applied Genetics* 122, 211–223.

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., *et al*. (2013) SLAF-seq: An efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLOS ONE* 8(3), e58700. DOI: 10.1371/journal.pone.0058700.

Sun, Y., Wang, J., Crouch, J.H. and Xu, Y. (2010) Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Molecular Breeding* 26, 493–511.

Tanksley, S.D. (1993) Mapping polygenes. *Annual Review of Genetics* 27, 205–233.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., *et al*. (2013a) QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal* 74, 174–183.

Takagi, H., Uemura, A., Yaegashi, H., Tamiru, M., Abe, A., *et al*. (2013b) MutMap-Gap: Whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. *New Phytologist* 200, 276–283.

Takagi, H., Tamiru, M., Abe, A., Yoshida, K., Uemura, A., *et al*. (2015) MutMap accelerates breeding of a salt-tolerant rice cultivar. *Nature Biotechnology* 33, 445–449.
Tan, L.B., Li, X.R., Liu, F.X., Sun, X.Y., Li, C.G., *et al*. (2008) Control of a key transition from prostrate to erect growth in rice domestication. *Nature Genetics* 40, 1360–1364.
Terao, T., Nagata, K., Morino, K. and Hirose, T. (2010) A gene controlling the number of primary rachis branches also controls the vascular bundle formation and hence is responsible to increase the harvest index and grain yield in rice. *Theoretical and Applied Genetics* 120, 875–893.
The 3,000 rice genome project (2014) The 3,000 rice genomes project. *Gigascience* 3, 7.
Thoday, J.M. (1961) Location of polygenes. *Nature* 191, 368–370.
Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., *et al*. (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28, 286–289.
Truntzler, M., Barriere, Y., Sawkins, M.C., Lespinase, D., Betran, J., *et al*. (2010) Meta-analysis of QTL involved in silage quality of maize and comparison with the position of candidate genes. *Theoretical and Applied Genetics* 121, 1468–1482.
Uga, Y., Sugimoto, K., Ogawa, S., Rane, J., Ishitani, M., *et al*. (2013) Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. *Nature Genetics* 45, 1097–1102.
Utz, H.F. and Melchinger, A.E. (1996) PLABQTL: A program for composite interval mapping of QTL. *Journal of Agricultural Genomics* 2, 1–6.
Veyrieras, J.-B., Goffinet, B. and Charcosset, A. (2007) MetaQTL: A package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics* 8, 49.
Vikram, P., Swamy, B.M., Dixit, S., Ahmed, H., Cruz, M.S., *et al*. (2012) Bulked segregant analysis: An effective approach for mapping consistent-effect drought grain yield QTLs in rice. *Field Crops Research* 134, 185–192.
Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., *et al*. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.
Wang, E.T., Wang, J.J., Zhu, X.D., Hao, W., Wang, L.Y., *et al*. (2008) Control of rice grain-filling and yield by a gene with a potential signature of domestication. *Nature Genetics* 40, 1370–1374.
Wang, S., Basten, C.J. and Zeng, Z.-B. (2012a) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm (accessed 10 October 2019).
Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., *et al*. (2012b) Control of grain size, shape and quality by OsSPL16 in rice. *Nature Genetics* 44, 950–954.
Wang, Y., Zhang, H., Xie, J., Guo, B., Yongxing Chen, Y., *et al*. (2018) Mapping stripe rust resistance genes by BSR-Seq: *YrMM58* and *YrHY1* on chromosome 2AS in Chinese wheat lines Mengmai 58 and Huaiyang 1 are *Yr17*. *The Crop Journal* 6, 91–98.
Wei, X.J., Xu, J.F., Guo, H.N., Jiang, L., Chen, S.H., *et al*. (2010) *DTH8* suppresses flowering in rice, influencing plant height and yield potential simultaneously. *Plant Physiology* 153, 1747–1758.
White, J.W., Andrade-Sanchez, P., Gore, M.A., Bronson, K.F., Coffelt, T.A., *et al*. (2012) Field-based phenomics for plant genetics research. *Field Crops Research* 133, 101–112.
Xue, W.Y., Xing, Y.Z., Weng, X.Y., Zhao, Y., Tang, W.J., *et al*. (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nature Genetics* 40, 761–767.
Yan, W.H., Wang, P., Chen, H.X., Zhou, H.J., Li, Q.P., *et al*. (2011) A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Molecular Plant* 4, 319–330.
Yang, J., Hu, C., Hu, H., Yu, R., Xia, Z., *et al*. (2008) QTL-Network: Mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24(5), 721–723.
Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the USA*, 90, 10972–10976.
Zeng, Z.-B. (1994) Precision mapping of quantitative trait loci. *Genetics* 136, 1457–468.
Zhang, X., Wang, W., Guo, N., Zhang, Y., Bu, Y., *et al*. (2018) Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genomics* 19, 226.
Zhang, Y., Liu, J., Xia, X. and He, Z. (2014) *TaGS-D1*, an ortholog of rice *OsGS3*, is associated with grain weight and grain length in common wheat. *Molecular Breeding* 34(3), 1097–1107. DOI: 10.1007/s11032-014-0102-7.
Zhou, C., Wang, P. and Xu, Y. (2016) Bulked segregants analysis in genetics, genomics and crop improvement. *Plant Biotechnology Journal* 14, 1941–1955.
Zhu, C., Gore, M., Buckler, E.S. and Yu, J. (2008) Status and prospects of association mapping in plants. *The Plant Genome* 1, 5–20.

9 Genotype–Environment Interaction and Stability Analyses: An Update

Manjit S. Kang*
Kansas State University, Manhattan, Kansas, USA

Introduction

Genotype–environment interaction (GEI) is an age-old, universal issue that relates to all living organisms (Kang, 1998). Genotypes and environments interact to produce an array of phenotypes. GEI is the variation caused by the joint effects of genotypes and environments (Dickerson, 1962). Baker (1988) defined GEI as the difference between the phenotypic value and the value expected from the corresponding genotypic and environmental values. When responses of two genotypes to different levels of environmental stress are compared, an interaction is described statistically as the failure of the two response curves to be parallel (Baker, 1988). Recently, de Leon *et al*. (2016) defined GEI as the differential sensitivity of certain genotypes to different environments. The GEI issue is not only important in plant-breeding programmes but also in animal-breeding programmes (Lin and Lin, 1994; Montaldo, 2001; Hayes *et al*., 2016).

Genotype-by-environment interaction must be distinguished from phenotypic plasticity. Schlichting (1986) defined phenotypic plasticity as the ability of an individual organism to alter its physiology/morphology in response to changes in environmental conditions. It describes the range of phenotypes produced by *a single genotype* in different environments (Grogan *et al*., 2016). The different phenotypes are referred to as *norms of reaction* (Redei, 1982; Kang, 1998). Phenotypic plasticity is a useful mechanism in plants, as being immobile, plants, unlike animals, cannot move away from a stressful situation but can adjust their performance through phenotypic plasticity. For GEI to occur, there must be at least two genotypes ($df = 1$) tested in at least two environments ($df = 1$), which will yield GEI with 1 *df* (Kang, 1998). van Eeuwijk *et al*. (2016) consider the GEI problem as the building of predictive models for genotype-specific reaction norms.

We should also distinguish between GEI and genotype–environment correlation (covariance) (GEC). GEC occurs when phenotypic and environmental effects are not independent. When additional agronomic inputs are given or are necessary for certain genotypes to do well, a genotype-by-environment correlation is created (Kang, 1998). Interaction or GEI occurs when the difference between the average phenotypic value for two genotypes changes in different environments, but GEC refers to the situation when particular genotypes tend to be associated with positive, and others with negative, environmental effects (Crow, 1986; Doolittle, 1987).

* Email: manjit675264@gmail.com

© CAB International, 2020.

The phenotypic variance (V_P) = V_G + V_E + V_{GE} + 2 Cov_{GE}, where V_G = genetic variance, V_E = environmental variance, V_{GE} = GEI variance, and Cov_{GE} = covariance between genotype and environment.

Note that $2Cov_{GE}$ entity is present in addition to V_{GE}. Crow explained the GEC as follows: Suppose a measurement z is the sum of two components, x and y. Thus, we write z = x + y. Then the variance of z is given by:

$$V_z = V_x + V_y + 2Cov_{xy}.$$

If x and y are independent, then $Cov_{xy} = 0$.

Terms 'genotyping' and 'phenotyping' are commonly used these days. A third term, 'envirotyping' has recently been added in the context of GEI (Cooper *et al.*, 2016; van Eeuwijk *et al.*, 2016; Xu, 2016). Xu (2016) has pointed out that to decipher environmental impacts on crop plants, 'envirotyping,' complements genotyping and phenotyping, and that it contributes to crop modelling and phenotype prediction through its functional components, including GEI, genes responsive to environmental signals, biotic and abiotic stresses, and integrative phenotyping. Cooper *et al.* (2016) have indicated high-throughput genotyping, phenotyping and envirotyping applied within plant breeding multi-environment trials (METs) provide the data foundations for selection and tackling GEI through whole-genome prediction. Pauli *et al.* (2016) have suggested that developments in envirotyping and crop growth modelling can provide a useful framework for understanding plant development.

Envirotyping allows researchers to apply real-world conditions when assessing the performance of crops and it has a wide range of applications, including the development of a four-dimensional profile for crop science, which would include a genotype, phenotype, envirotype and time (https://www.cimmyt.org/publications/new-publications-new-environmental-analysis-method-improves-crop-adaptation-to-climate-change/; accessed 1 June 2019).

Crossover and Non-crossover Interactions

GEI can be grouped into two broad categories: crossover and non-crossover interactions. GEI is important only if genotypes switch ranks from one environment to another (Haldane, 1947). A brief discussion of each follows.

Crossover or qualitative interaction

The differential response of cultivars to diverse environments is referred to as a crossover interaction when cultivar ranks change from one environment to another. A main feature of crossover interaction is intersecting lines in a graphical representation. If the lines do not intersect, there is no crossover interaction (Kang, 1998).

Variation among genotypes in phenotypic sensitivity to the environment (GEI) may necessitate the development of locally adapted varieties (Falconer, 1952). If no one genotype has superiority in all situations, GEI indicates the potential for genetic differentiation of populations under prolonged selection in different environments (Via, 1984).

In crop breeding, the crossover interaction is more important and problematic than non-crossover interaction (Baker, 1990). As the presence of a crossover interaction has strong implications for breeding for specific adaptation, it is important to assess the frequency of crossover interactions (Singh *et al.*, 1999). According to Gregorius and Namkoong (1986), crossover interaction is not only non-additive in nature but also non-separable. Lack of crossover interaction for quantitative trait loci (QTL) even in the presence of significant GEI has been reported (Lee, 1995; Beavis and Keim, 1996). The reader may refer to Beavis and Keim (1996), Cornelius *et al.* (1996), Crossa *et al.* (1996) and Singh *et al.* (1999) for further discussion on crossover and non-crossover interactions. Lee *et al.* (2016) have suggested that crossover GEI effects occur in maize when genotypes respond differentially to environmental factors that impact development, and non-crossover GEI effects occur when genotypes respond differentially to growth-related environmental parameters.

Non-crossover or quantitative interaction

These interactions represent changes in magnitude of genotype performance (quantitative),

but rank order of genotypes across environments remains unchanged. Non-crossover interactions may mean that genotypes are genetically heterogeneous but test environments are more or less homogeneous or that genotypes are genetically homogeneous but environments are heterogeneous. All identical genotypes grown in constant (ideal) environments should perform consistently. Any departure from the ideal environment leads to GEI.

Importance of GEI

Thus far, agricultural production has kept pace with the world's population growth mainly because of the innovative ideas and efforts of agricultural researchers. The world population, currently 7.7 thousand million, is expected to reach 9.8 thousand million by 2050 (https://www.worldometers.info/world-population/; accessed 9 June 2019). The key to enhancing agricultural production is increased efficiency in the utilization of resources (increased productivity per hectare and per dollar) and this includes a better understanding of GEI and ways of exploiting it. The importance of GEI is highlighted by Gauch and Zobel (1996):

> Were there no interaction, a single variety of wheat (*Triticum aestivum* L.) or corn (*Zea mays* L.) or any other crop would yield the most the world over, and furthermore the variety trial need be conducted at only one location to provide universal results. And were there no noise, experimental results would be exact, identifying the best variety without error, and there would be no need for replication. So, one replicate at one location would identify that one best wheat variety that flourishes worldwide.
>
> (Gauch and Zobel, 1996, p.87)

The importance of GEI can be seen from the relative contributions of new cultivars and improved management to yield increases from direct comparisons of yields of old and new varieties in a single trial (Silvey, 1981). Genetic improvements have been estimated to account for about 50% of the total gains in yield per unit area for major crops during the past 60–70 years (Silvey, 1981; Simmonds, 1981; Duvick, 1992, 1996). The remainder of the yield gain is attributable to improved management and cultural practices. Barley yield data from the UK (1946–1977: mean yield for 1946 = 2.3 t ha^{-1} and for 1977 = 3.9 t ha^{-1}) indicated that the environmental contribution was 10–30% and the genetic contribution 30–60%; the remainder 25–45% of the yield gain was attributed to GEI (Simmonds, 1981). For wheat for the same period (1946–1977: mean yield for 1946 = 2.4 t ha^{-1} and for 1977 = 4.7 t ha^{-1}), yield gain was attributed as follows: 40–60% to the environment (E), 25–40% to the genotype (G) and 15–25% to GEI (Simmonds, 1981). The GEI confounds precise partitioning of the contributions of improved cultivars and improved environment/technology to yield (Silvey, 1981). Thus, the combined contributions of G and genotype–environment (GE) effects can be substantial (40–60% wheat and 70–90% in barley).

GEI occurs during, and has an impact on all stages of a breeding programme and has enormous implications for the allocation of resources. A large GEI could mean the establishment of two fully fledged breeding stations in a region, instead of one, thus requiring increased input of resources (manpower, land and money) (see Kang, 1998).

Heritability of a trait plays a key role in determining genetic advance from selection. As a component of the total phenotypic variance (the denominator in any heritability equation), GEI affects heritability negatively. The larger the GEI component, the smaller the heritability estimate; thus, progress from selection would be limited.

A large GEI reflects the need for testing cultivars in numerous environments (locations and/or years) to obtain reliable results. If the weather patterns and/or management practices differ in target areas, testing must be done at several sites representative of the target areas.

Kang (1993a) discussed the disadvantages of discarding genotypes evaluated in only one environment in early stages of a breeding programme. The discarded genotypes might have the potential to do well at another location or in another year. Thus, some potentially useful genes could be 'lost' because of limited testing. An example from six-row barley illustrates this point well. A total of 288 barley lines were evaluated in the Magreb countries and in International Center for Agricultural Research in Dry Areas (ICARDA)'s yield trials at three locations (Ceccarelli *et al.*, 1994). Of the 103 lines selected at ICARDA and 154 lines at the Magreb, only 49 were selected at both locations.

Performance evaluation is the second important component of a breeding programme. Testing done in one environment provides only limited information. Multi-environment testing provides additional useful information, e.g. a GEI component can be estimated. In addition, multi-environment testing yields better estimates of variance components and heritability. Therefore, GEI need not be perceived only as a problem.

As the magnitude of a significant interaction between two factors increases, the usefulness and reliability of the main effects are correspondingly decreased. Because GEI reduces the correlation between phenotypic and genotypic values, the difficulty in identifying truly superior genotypes across environments is magnified.

Obviously, the cost of cultivar evaluation increases as additional testing is carried out. However, with additional test environments, a breeder/agronomist can identify cultivars with specific adaptation as well as those with broad adaptation, which will not be possible from testing in a single environment. Broad adaptation provides stability against the variability inherent in an ecosystem, but specific adaptations may provide a significant yield advantage in particular environments (Wade *et al.*, 1999). Multi-environment testing makes it possible to identify cultivars that perform consistently from year to year (small temporal variability) and those that perform consistently from location to location (small spatial variability). Temporal stability is desired by and is beneficial to growers, whereas spatial stability is beneficial to seed companies and breeders. Stability of performance can be ascertained via stability statistics (Lin *et al.*, 1986; Kang, 1990; Kang and Gauch, 1996).

According to Kleinknecht *et al.* (2016), variety testing involves the challenge to design multi-environmental trials in several years and locations. They indicated that several variables influenced the varietal performance and provided a simulation-based approach using Statistical Analysis System (SAS) to vary those variables and to allow a comparison of different scenarios for the design of a series of trials regarding selection gain.

Achievements

The GEI issue received focused attention in 1990 when an international symposium on 'Genotype-by-Environment Interaction and Plant Breeding' was held on 12 and 13 February at the Louisiana State University campus in Baton Rouge (Kang, 1990). The various GEI issues have come to the forefront in many breeding programmes throughout the world. Reviews and extensive bibliographies (Aastveit and Mejza, 1992; Annicchiarico and Perenzin, 1994; Denis and Gower, 1996; Denis *et al.*, 1996a; Kang, 1998; Piepho, 1998), conference/symposia proceedings (Rao *et al.*, 1988, 1993; Cooper and Hammer, 1996; Zavala-Garcia and Treviño-Hernández, 2000) and books (Gauch, 1992; Prabhakaran and Jain, 1994; Hildebrand and Russell, 1996; Kang and Gauch, 1996; Hoffmann and Parsons, 1997; Basford and Tukey, 1999; Hall, 2001) have since been published. Recognizing the importance of GEI in plant breeding, the Crop Science Society of America organized a symposium on this issue and published papers in *Crop Science* volume 56. Several of those articles have been reviewed in this chapter.

GEI presents many challenges for breeders and has significant implications in both applied plant- and animal-breeding programmes. The breeder is faced with developing separate populations for each site type where genotypic rankings drastically change and/or is faced with selecting genotypes that generally perform well across many sites (McKeand *et al.*, 1990). Gains are expected to be greater with the first approach, but costs would also likely be higher; the second approach, while less expensive, yields smaller gains. Denis and Gower (1996) suggested that plant breeders should consider GEI to avoid missing a variety that performed, on average, poorly but did well when grown in specific environments or selecting a variety that, on average, performed well but did poorly when grown in a particular environment.

Denis *et al.* (1996b) presented a number of models that can account for heteroscedasticity in GE tables. They presented a general scheme for describing heteroscedasticity with a reduced number of parameters using the mixed-model framework, which allows new parsimonious models.

Since the 1970s, various attempts have been made to jointly capture the effects of G and GEI. Simultaneous selection for yield and stability of performance is an important consideration in breeding programmes. No methods

developed so far have been universally adopted. Flores *et al*. (1998) compared 22 univariate and multivariate methods to analyse GEI. These 22 methods were classified into three main groups (Flores *et al*., 1998): Group 1 statistics were mostly associated with yield level and showed little or no correlation with stability parameters; Group 2 statistics considered both yield and stability of performance simultaneously to reduce the effect of GEI; and Group 3 statistics emphasized only stability. Group 1 included YIELD, PI, UPGMA, FOXRANK and FOXROS; Group 2 included S6O, PPCC, STAR, AMMI and KANG; and Group 3 included TAI, LIN, CA, SHUKLA and EBRAS.

Hussein *et al*. (2000) provided a comprehensive statistical analysis system (SAS) program for computing univariate and multivariate stability statistics for balanced data. Their program provides estimates of more than 15 stability-related statistics.

Path coefficient analysis has been effectively used to investigate GEI in potato by Tai and Coleman (1999). The path analysis has not found much favour with most researchers. Nevertheless, Tai has expounded on the merits of this method (Tai, 1990).

Piepho (2000b) proposed a mixed-model method to detect QTL with significant mean effect across environments and to characterize the stability of effects across multiple environments. He treated environment main effects as random, which meant that both environmental main effects and QTL-by-environment interaction (QEI) effects could be regarded as random.

Biadditive factorial regression models, which encompass both factorial regression and biadditive (additive main effect and multiplicative interaction [AMMI]) models, have also been evaluated (Brancourt-Hulmel *et al*., 2000). The biadditive factorial regression models involved environmental covariates related to each deviation and included environmental main effect, sum of water deficits, an indicator of nitrogen stress, sum of daily radiation, high temperature, pressure of powdery mildew and lodging (Brancourt-Hulmel *et al*., 2000). The models explained about 75% of the interaction sum of squares. The biadditive factorial biplot provided relevant information about the interaction of the genotypes with respect to environmental covariates. Paderewski *et al*. (2016) recently used wheat MET data from Poland and extended AMMI analysis from the customary two-way G×E datasets to four-way datasets, such as a genotype-by-location-by-management-by-year (GLMY) dataset.

Piepho *et al*. (2016) estimated variances of genotype × location and genotype × location × year interactions from wheat trials to assess stability. They found substantial differences in stability among the 16 genotypes that were evaluated.

AMMI has been used extensively for analysing METs for two purposes, i.e. understanding complex GEIs and increasing accuracy (Gauch and Moran, 2019). In relation to AMMI, Gauch and Furnas (1991) reported the development of MATMODEL, which enabled researchers to (i) increase the accuracy of yield estimates; (ii) improve selections; (iii) impute missing data; (iv) model and understand the genotypes, environments and interaction, particularly with a biplot graph; and (v) design flexible and efficient experiments. Gauch and Moran (2019) have recently reported new software called AMMISOFT, which facilitates AMMI analyses to help accelerate crop improvement.

The biplot method originated with Gabriel (1971). Others have used this method in describing GEI. The versatility of the GGE (G = genotype effect and GE = genotype–environment effect) biplot was elucidated by Yan *et al*. (2000). In addition to dissecting GEIs, GGE Biplot helps analyse genotype-by-trait data, genotype-by-marker data, and diallel cross data (Yan *et al*., 2000, 2001; Yan, 2001; Yan and Hunt, 2001, 2002; Yan and Rajcan, 2002). Glaz and Kang (2008) determined the best test locations from sugarcane trial data using GGE biplot analysis. These aspects make the GGE biplot a most comprehensive tool in quantitative genetics and plant breeding (see Yan and Hunt, Chapter 10, this volume). The GGE Biplot methodology and its applications have been described in detail by Yan and Kang (2003) and Yan (2014).

Causes of Genotype–Environment Interaction

To be able to understand GEI and utilize it effectively in breeding programmes, information is needed on the factors responsible for the differential response of genotypes to variable

environments. A factor may be present at optimal, suboptimal or super-optimal levels. When present at a level other than optimal, it represents a stress. According to Baker (1988), differences in the rate of increase in response of genotypes at suboptimal levels would reflect differences in efficiency, and differences in the rate of decrease at super-optimal levels would reflect differences in tolerance. Without the presence of stresses, genotype attributes, such as efficiency and tolerance, cannot be identified and investigated. In this section, the effects of environmental stress on the plant genome in general and biotic and abiotic factors that may be responsible for GEI are considered.

Environmental effect on the genome

An understanding of plant stress responses is essential because of predicted global environmental changes and their impact on the production of food and fibre. Stress is a physiological response to an adverse environmental factor(s). Plants respond to a variety of environmental cues: nutrients, toxic elements and salts in the soil solution, gases in the atmosphere, light of different wavelengths, mechanical stimuli, gravity, wounding, pests, pathogens and symbionts (Crispeels, 1994). Plants have incorporated a variety of environmental signals into their developmental pathways that have provided for their wide range of adaptive capacities across time (Scandalios, 1990).

Environmental stresses have been shown to elicit specific responses at the DNA level in a number of organisms. A differentiated cell expresses an array of genes required for its stable functioning and metabolic roles (Scandalios, 1990). In response to severe environmental changes, a genome can respond by selectively regulating (increasing or decreasing) the expression of specific genes.

Interspecific variation in DNA amounts is correlated with various quantitative properties of cells, and these may secondarily affect the quantitative characters of the whole plant (Bachmann *et al.*, 1985; Cavalier-Smith, 1985a,b; Bennett, 1987). Highly significant differences of up to 32% in DNA content were found in meristems of seedlings from 35 natural populations of hexaploid *Festuca arundinacea* (Ceccarelli *et al.*, 1992). In cultivated maize, variation in genome size has been reported to be as high as 38.8% (Laurie and Bennett, 1985; Rayburn *et al.*, 1985). Maize lines from higher latitudes of North America had significantly lower nuclear DNA amounts than those from lower latitudes (Rayburn *et al.*, 1985). Rayburn and Auger (1990) determined the nuclear DNA content of 12 southwestern US maize populations collected at various altitudes and observed a significant positive correlation between genome size and altitude. Higher amounts of DNA at higher elevation have also been found in teosinte (Laurie and Bennett, 1985).

Herrera-Estrella and Simpson (1990) investigated the influences of environmental factors on genes involved in photosynthesis. The mechanism of regulation may vary from one species to another (Herrera-Estrella and Simpson, 1990).

Biotic stresses

Biotic stress factors are a major limitation to plant productivity and a dominant element in plant ecology and evolution (Higley *et al.*, 1993). Biotic stresses and interactions among them and/or with abiotic factors remain poorly understood; however, they have significant relevance to GEI in plants.

Plants may respond to pathogen infection by inducing a long-lasting, broad-spectrum, systemic resistance to subsequent infections (Ryals *et al.*, 1994). Induced disease resistance has been referred to as physiological acquired immunity, induced resistance or systemic acquired resistance (SAR). Differences in insect and disease resistance among genotypes can be associated with stable or unstable performance (Baker, 1990).

Abiotic stresses

The major abiotic stresses are atmospheric pollutants, soil stresses (salinity, acidity, and mineral toxicity and deficiency), temperature (heat and cold), water (drought and flooding) and tillage operations (Blum, 1988; Clark and Duncan, 1993; Specht and Laing, 1993; Unsworth and Fuhrer, 1993). Genetic variation exists for plant

responses to the above stress factors. Breeding for tolerance to air pollutants has considerable potential (Unsworth and Fuhrer, 1993).

With stress caused by suboptimal levels of water, nutrients and solar radiation, it should be possible to identify genotypes that are efficient or inefficient in using the respective resource. Woodend and Glass (1993) demonstrated the presence of GEI for potassium-use efficiency in wheat.

Responses to temperature

Rapid temperature changes, particularly those towards the upper end of the adaptation range for individual plant species, can produce dramatic changes in the pattern of gene expression. Heat-shock responses are plants' protective measures against potentially lethal, rapid-rate, upward departures from the optimal temperature (Pollack *et al.*, 1993). Tolerance of protein synthesis and seedling growth to a previously lethal high temperature can be induced by prior short exposure to a sublethal high temperature that triggers the synthesis of a specific set of proteins – the heat-shock proteins (HSPs) – via mRNA that is newly transcribed in response to high temperature. In the meantime, the synthesis of normal cellular proteins is reduced or shut down. This process is detectable within minutes of the onset of stress (Ougham and Howarth, 1988). HSPs are induced at different temperatures in different species. The rule of thumb is that temperature must be ~10°C higher than the optimal temperature for a particular species.

Oxidative stress

A common feature of different stress factors is an increased production of reactive oxygen species in plant tissues, but their mode of action varies depending on whether oxidants are generated outside (e.g. by oxidizing air pollutants) or inside a plant cell (e.g. high radiation, low temperatures or nutrient deficiency) (Polle and Rennenberg, 1993). It is important to understand both the mode of action of different stress factors and the critical physiological properties that limit ameliorative mechanisms at the subcellular level (Polle and Rennenberg, 1993).

Scandalios (1990) summarized plant responses to environmental stress, pointing out that activated oxygen species (endogenous – by-products of normal metabolism – and exogenous – triggered by environmental factors) were highly reactive molecules capable of causing extensive damage to plant cells. The effects of oxidative stress can range from simple inhibition of enzyme function to the production of random lesions in proteins and nucleic acids and the peroxidation of membrane lipids. Loss of membrane integrity can cause decreased mitochondrial and chloroplast functions, which, in turn, can lower the plant's ability to fix carbon and to properly utilize the resulting products (Scandalios, 1990). This decrease in metabolic efficiency results in reduced yield.

Reckling *et al.* (2015) studied the causes of low stability of two grain legume species and reported the low yield stability might be associated with higher pest, disease and weed infestations, and physiological aspects. They conjectured that growth habit of grain legumes not being determinate could have reduced their ability to cope with the fluctuating water availability that was typical for the studied site.

How to Deal with GEI

The presence of crossover interactions has important implications for breeding strategies that aim to improve either broad or specific adaptation or some combination of both components of adaptation (Cooper *et al.*, 1999). Eisemann *et al.* (1990) listed three ways of dealing with GEI in a breeding programme: (i) ignore them, i.e. use genotypic means across environments even when GEI exists; (ii) avoid them; or (iii) exploit them. Interactions should not be ignored when they are significant and of the crossover type.

The second way of dealing with these interactions, i.e. avoiding them, involves minimizing the impact of significant interactions. One approach is to group similar environments (forming mega-environments) via a cluster analysis. With environments being more or less homogeneous, genotypes evaluated in them would not be expected to show crossover interactions. By clustering environments, potentially useful information may be lost. International research centres, such as the International Maize and Wheat Improvement Center (CIMMYT), aim to

identify maize and wheat genotypes with broad adaptation (i.e. stable performance across diverse environments) at many international sites. Such an objective cannot be achieved by grouping (clustering) test environments.

The third approach encompasses stability of performance across diverse environments by analysing and interpreting genotypic and environmental differences. This approach allows researchers to select genotypes with consistent performance, identify the causes of GEI and provide the opportunity to correct the problem. When the cause for the unstable performance of a genotype is known, either the genotype can be improved by genetic means or a proper environment (inputs and management) can be provided to enhance its productivity.

Yan (2016) pointed out two viable options to deal with GEI: to utilize it or to avoid it, depending on whether it is repeatable. He suggested that repeatable GEI can be selected for (utilized), whereas non-repeatable GEI must be selected against (avoided). Utilization of GEI involves: (i) identifying repeatable GEI; (ii) dividing the target region into subregions or mega-environments based on the repeatable GEI pattern; and (iii) selecting within mega-environments (Yan, 2016). GEI within mega-environments should be avoided as it is non-repeatable. Genotype selection should be based on both high mean performance and high stability (Kang, 1993b, 1998; Yan, 2016).

Lado *et al.* (2016) compared strategies to exploit GEI in genomic selection (GS) using mixed models. They specifically compared strategies to predict new genotypes by borrowing information from other environments modelling the correlation matrix across environments and to design sets of environments aiming for low GEI to predict genomic performance in new environments. A genotype that performs consistently (high-yielding) across many environments would possibly possess broad-based, durable resistances/tolerances to the biotic and abiotic environmental factors that it encountered during development. The more the breeders know about the crop environment, the better job they can do of judiciously targeting appropriate cultivars to production environments.

In the next section, the concepts of stability are presented. A method for identifying stable genotypes and environmental factors that may be responsible for stable or unstable performance is also given.

Concepts of stability

Stability is a central keyword for plant breeders analysing GEI data. A simple corresponding statistical term is 'dispersion around a central value' (Denis *et al.*, 1996a). There are two concepts of stability: static and dynamic. The static concept means that a genotype has a stable performance across environments and there is no among-environment variance. This would mean that a genotype would not respond to high levels of inputs, such as fertilizers. This type of stability would not be beneficial for the farmer, and it has been referred to as the biological concept of stability (Becker, 1981), which is equivalent to Lin *et al.*'s (1986) type 1 stability. In type 1 stability, a genotype is regarded as stable if its among-environment variance is small.

The dynamic concept means that a genotype has a stable performance, but, for each environment, its performance corresponds to the estimated level or predicted level. There would be agreement between the estimated or predicted level and the level of actual performance (Becker and Leon, 1988). This concept has been referred to as the agronomic concept (Becker, 1981), which is equivalent to Lin *et al.*'s (1986) type 2 stability. In type 2, a genotype is regarded as stable if its response to environments is parallel to the mean response of all genotypes in a test.

Lin *et al.* (1986) defined four groups of stability statistics. Group A is based on deviation from the average genotype effect (DG), group B on the GEI term (GEI), and groups C and D on either DG or GEI. The formulae of groups A and B represent sums of squares and those of groups C and D represent a regression coefficient or deviation from regression. They integrated type 1, type 2 and type 3 stabilities with the four groups: group A was regarded as type 1, groups B and C as type 2, and group D as type 3 stability. In type 3 stability, a genotype is regarded as stable if the residual mean square from the regression model on the environmental index is small (Lin *et al.*, 1986). Lin and Binns (1988) proposed the type 4 stability concept on the basis of predictable and unpredictable non-genetic variation: the

predictable component is related to locations and the unpredictable component is related to years. Lin and Binns (1988) suggested the use of a regression approach for the predictable portion and the mean square for years within locations for each genotype as a measure of the unpredictable variation. The latter was called the type 4 stability statistic.

Stability statistics

Plant breeding can exploit wide adaptation by selecting genotypes that yield well across large geographical areas or mega-environments (Witcombe, 2001). Mega-environments are broad (frequently discontinuous transcontinental) areas that are characterized by similar biotic and abiotic stresses, cropping-system requirements and consumer preferences (Witcombe, 2001). Several methods have been developed to analyse GEI and to select genotypes that perform consistently across many environments (Lin *et al.*, 1986; Becker and Leon, 1988; Kang, 1990, 1998; Kang and Gauch, 1996; Weber *et al.*, 1996). The earliest approach was the linear regression analysis (Mooers, 1921; Yates and Cochran, 1938). Finlay and Wilkinson (1963), Eberhart and Russell (1966), and Tai (1971) popularized variations of the regression approach, assuming an expected linear response of yield to environments. The merits and demerits of several methods were discussed by Kang and Miller (1984). Kang *et al.* (1987b) concluded that Shukla's (1972) stability variance and Wricke's (1962) ecovalence were equivalent methods and they ranked genotypes identically for stability (rank correlation coefficient = 1.00). These types of measures are useful to breeders and agronomists, as they provide the contribution of each genotype to total GEI. They can also be used to evaluate testing locations by identifying those locations with a similar GEI pattern (Glaz *et al.*, 1985). Other statistical methods that have received significant attention are pattern analysis (DeLacy *et al.*, 1996), the AMMI model (Gauch and Zobel, 1996), the shifted multiplicative model (SHMM) (Cornelius *et al.*, 1996; Crossa *et al.*, 1996), the non-parametric methods of Hühn (1996), which are based on cultivar ranks, the probability of outperforming a check (Eskridge, 1996) and Kang's rank-sum method (Kang, 1988, 1993b). The methods of Hühn (1996) and Kang (1988, 1993b) integrate yield and stability into one statistic that can be used as a selection criterion.

Dashiell *et al.* (1994) evaluated the usefulness of several stability statistics for simultaneously selecting for high yield and stability of performance in soybean. Fernandez (1991) also evaluated stability statistics for similar purposes. Flores *et al.* (1998) and Hussein *et al.* (2000) conducted comparative evaluations of 22 and 15 stability statistics/methods, respectively.

Simultaneous selection for yield and stability

Growers would prefer to use a high-yielding cultivar that performs consistently from year to year (temporal adaptation) and might be willing to sacrifice some yield if they are guaranteed, to some extent, that a cultivar would produce consistently from year to year (Kang *et al.*, 1991). Kang (1993b) discussed the motivation for emphasizing stability in the selection process. He enumerated the consequences to growers of researchers' committing type I (rejecting the null hypothesis when it is true) and type II errors (accepting the null hypothesis when it is false) relative to selection on the basis of yield alone (conventional method [CM]) and that on the basis of yield and stability. Simultaneous selection for yield and stability reduces the probability of committing type II errors (probability = β). Generally, type II errors constitute the most serious risk for growers (Glaz and Dean, 1988; Johnson *et al.*, 1992). The combined rate of committing a type II error for simultaneous selection for yield and stability will be the product of β for comparisons of overall yield mean and β for comparisons of GEI means.

Several methods of simultaneous selection for yield and stability and relationships among them were discussed by Kang and Pham (1991). The development and use of the yield–stability statistic (YS_i) demonstrated the significance and rationale of incorporating stability in selecting genotypes tested across a range of environments (Kang, 1993b). A QBASIC computer program (STABLE) for calculating this statistic has been

developed and is available free of charge (Kang and Magari, 1995).

The stability component in YS_i is based on Shukla's (1972) stability-variance statistic (σ_i^2). Shukla (1972) partitioned GEI into components, one corresponding to each genotype, and referred to it as stability variance. Lin *et al.* (1986) classified σ_i^2 as type 2 stability, meaning that it was a relative measure dependent on genotypes included in a particular test. Pazdernik *et al.* (1997) analysed soybean seed yield, protein and oil concentrations and stability statistics. They concluded that Hühn's rank-based S_i^1 and S_i^2 statistics and Kang's YS_i statistic could be used by breeders to select parents to improve protein concentration and stability by combining stable high-yielding lines with stable high-protein lines. They further suggested that the same statistics could be used by consultants and variety-testing personnel to aid in making recommendations to soybean producers.

Covariates and stability

Yield stability or GEI for yield is a complex issue. Yield stability depends on plant characteristics, such as resistance to pests and tolerance to environmental stress factors. By determining factors responsible for GEI or stability/instability, breeders can improve cultivar stability. If instability was caused by susceptibility to a disease, breeding for resistance to that disease should reduce losses in disease-inducing environments and increase genotype stability.

It is important to know the underlying causes of GEI (Kang, 1998; Haji and Hunt, 1999). An observational description of GEI is not very useful unless one knows the elements that cause the environmental differentiation (Federer and Scully, 1993). The use of environmental variables as covariates was suggested and/or employed by several researchers (Freeman and Perkins, 1971; Hardwick and Wood, 1972; Shukla, 1972; Wood, 1976; Kang and Gorman, 1989; van Eeuwijk *et al.*, 1996; Piepho *et al.*, 1998). Individual components of the environment (rainfall, temperature, fertility, etc.), used as covariates in explaining GEI, can greatly increase the reliability of predictions relative to cultivar performance. Environmental characterization can be achieved directly, by measuring environmental variables, which can be physical, biological or nutritional, or indirectly, by measuring plant responses to capture the influence of environmental conditions on plant performance (Brancourt-Hulmel *et al.*, 2000). Winter-wheat data from Ontario revealed that January temperatures, together with moisture supply before anthesis, were associated with some of the GEIs (Haji and Hunt, 1999). Lee *et al.* (2016) have tried to relate year-to-year variation in thermal time, solar radiation and soil available moisture to GEI in maize.

A fertility score was used as an environmental covariate in Germany (Piepho, 2000a). This score, ranging between 0 and 100, incorporates several variables, including soil type and the geological age of parent material. Piepho (2000a), who provided confidence limits for estimated risks, argued that, if yield depends on environmental covariates, risk for a specific environment can be estimated on the basis of covariate information, thus yielding a more specific risk assessment.

Methods of assessing the contributions of weather variables and other factors (covariates) that contribute to GEI are available (Shukla, 1972; Denis, 1988; van Eeuwijk *et al.*, 1996; Magari *et al.*, 1997). Contributions of different environmental variables to GEI have been reported by several researchers (Saeed and Francis, 1984; Gorman *et al.*, 1989; Kang and Gorman, 1989; Kang *et al.*, 1989; Rameau and Denis, 1992; Magari *et al.*, 1997; Vargas *et al.*, 2001; Yan and Hunt, 2001).

In the following linear model, GEI is explained with a covariate, as shown by Shukla (1972):

$$Y_{ijk} = \mu + \alpha_{i+}\theta_{ij} + \beta_k + b_k z_i + \varepsilon_{ijk} \qquad \text{(Eqn 9.1)}$$

where Y_{ijk} = observed trait value, μ = grand mean, α_i = environmental effect, θ_{ij} = blocks within environments effect, β_k = cultivar effect, b_k = regression coefficient of the kth genotype's yield in different environments, z_i = an environmental covariate and ε_{ijk} = experimental error.

When a number of environmental variables are considered, a combination of two or more variables would remove more heterogeneity from GEI than individual variables do. Methods developed by van Eeuwijk *et al.* (1996) may be helpful for this purpose. Magari *et al.* (1997) identified

precipitation as the single most important environmental factor that contributed to GEI for ear-moisture loss rate in maize. They identified precipitation + growing degree-days from planting to black-layer maturity (GDD-BL) and relative humidity + GDD-BL as the two-factor combinations that explained larger amounts of GEI compared with other combinations.

Vargas *et al.* (2001) found the most important variables that explained nitrogen (N)–year interaction to be minimum temperature in January–March and maximum temperature in April. Evaporation rates for December and April were important covariates for describing tillage–year and summer crop–year interactions, whereas precipitation in December and sun hours in February explained year–manure interaction (Vargas *et al.*, 2001).

Stability variance for unbalanced data

Plant breeders often deal with unbalanced data. Searle (1987) classified unbalancedness as planned unbalanced data and missing observations. Both categories of unbalancedness may occur, but planned unbalancedness (a situation when, for different reasons, one does not have data for all genotypes in all environments) is more difficult to handle. Researchers have used different approaches for studying GEI in unbalanced data (Freeman, 1975; Pedersen *et al.*, 1978; Zhang and Geng, 1986; Gauch and Zobel, 1990; Rameau and Denis, 1992; Piepho, 1994). Usually environmental effects are regarded as random and cultivar effects as fixed. Inference on random effects using least squares, in the case of unbalanced data, is not appropriate because information on variation among random effects is not incorporated (Searle, 1987). For this reason, mixed-model equations (MMEs) are recommended (Henderson, 1975).

The values of Shukla's (1972) σ_i^2 can be negative because they are calculated as the differences of two statistically dependent sums of squares, which is a negative feature of this approach. Computation of σ_i^2 is impossible from unbalanced data, but genotype$_k$–environment variance components ($\sigma^2_{g(k)e}$) can be estimated using the maximum likelihood approach. The general linear model for randomized complete-block-design experiments conducted in different environments is:

$$Y_{ijk} = \mu + \alpha_i + \theta_{ij} + \beta_k + \gamma_{ik} + \varepsilon_{ijk} \quad \text{(Eqn 9.2)}$$

Using matrix notation, Eqn 9.2 can be written as:

$$y = 1\mu + X\beta + W\alpha + U\theta + \Sigma_k Z_k a_k + \varepsilon \quad \text{(Eqn 9.3)}$$

where y = vector of observed yield data, 1 = vector of ones, X = design matrix for fixed effects (genotypes), β = vector of genotype effects, W and α are, respectively, a design matrix for and a vector of environmental effects, U and θ are, respectively, a design matrix for and a vector of replications within environment effects, Zk and ak are, respectively, a design matrix for and a vector of GEI effects, and ε is the vector of residuals. Equation (9.3) can be solved using Henderson's (1975) MME. The levels of random factors are generally assumed to be independent.

The restricted maximum-likelihood (REML) methodology is generally preferred to maximum-likelihood estimates because it considers the degrees of freedom for fixed effects for calculating error. The calculation of REML stability variances for unbalanced data allows one to obtain a reliable estimate of stability parameters and overcomes the difficulties of manipulating unbalanced data (Kang and Magari, 1996).

Testing and Breeding Strategies

The best approach for breeders and geneticists would be to understand the nature and causes of GEI and to try to minimize its deleterious implications and exploit its beneficial potential through appropriate breeding, genetic and statistical methodologies (Kang and Gauch, 1996). Appropriate analyses of data can provide an opportunity for exploiting GEI using applied analytical methods, such as AMMI, GGE biplots, the use of climatic factors in explaining GEI, and the evaluation of risk of production and the optimal allocation of land resources to various genotypes for selection in heterogeneous environments (Singh *et al.*, 1999). Some of the important strategies for accomplishing this are outlined in the following paragraphs.

Breeding for resistance/tolerance to stresses

Resistance or tolerance to any type of stress, biotic or abiotic, is essential for stable performance (Khush, 1993; Duvick, 1996). Sources of increased crop productivity include enhanced yield potential, heterosis, modified plant types, improved yield stability, gene pyramiding and exotic and transgenic germplasm (Khush, 1993). It is important to identify the factor(s) that are responsible for GEI. If interaction is caused by European corn-borer (ECB) damage, a gene conferring resistance to ECB could be inserted into one of the two inbred parents of the susceptible hybrid genotype.

Brancourt-Hulmel (1999) used crop diagnosis with the analysis of interaction by factorial regression in wheat. She provided an agronomic explanation of GEI and defined responses or parameters for each genotype and each environment. Earliness at heading, susceptibility to powdery mildew, and susceptibility to lodging were the major factors responsible for GEI. In the same study (Brancourt-Hulmel, 1999), factorial regression revealed that water deficits during the formation of grain number and N level also were associated with GEI.

To alleviate GEI concerns caused by stresses, breeders need to know as much about the various characteristics of genotypes as possible. They also need to characterize environments as fully as possible. Knowledge of soil characteristics and ranges of weather variables and stresses that plant materials will be exposed to is a prerequisite to exploiting the beneficial potentials of the genotypes and environments and to targeting appropriate cultivars to specific environments.

Economically important traits in crops are generally quantitative in nature. For improving quantitative traits, breeders need to know what genetic factors are involved, where they are located on chromosomes and what type of inheritance they exhibit. Recent advances in molecular genetics have provided some of the best tools for obtaining insights into the molecular mechanisms associated with GEI. Molecular markers, such as restriction fragment-length polymorphisms (RFLPs), can be employed to find genomic regions with stable responses. Molecular markers have paved the way for investigating the QTL-by-environment interaction (QEI) (Beavis and Keim, 1996), which will ultimately provide a better genetic understanding of this phenomenon and its possible regulation/exploitation. Regions of plant genomes that provide stable responses across diverse environments can be identified by determining the linkage of QTL to RFLPs, which should make it possible for breeders to manipulate QTL in the same fashion as single genes that control qualitative traits. Wang *et al.* (1999) reported a new methodology based on mixed linear models to map QTL with digenic epistasis and QEIs. Reliable estimates of QTL main effects (additive and epistatic effects) can be obtained with the maximum-likelihood estimation method, and QEI effects (additive–environment interaction and epistatic effects–environment interaction) can be obtained with the best linear unbiased prediction (BLUP) method (Wang *et al.*, 1999).

Rodrigues (2018) has presented an overview of statistical methods and models commonly used to detect and to understand GEI and QEI, which ranged from the simple joint regression model to complex eco-physiological genotype-to-phenotype simulation models. Researchers can benefit from knowledge of these interactions in selecting better genotypes across different environmental conditions, and consequently to improve crops in developed and developing countries (Rodrigues, 2018). Malosetti *et al.* (2016) suggested that genomic prediction (GP) can assist selection decisions by combining incomplete phenotypic information across multiple environments with dense sets of markers.

It is highly desirable to identify QTL for a complex trait (say, high yield) that is expressed in a number of environments. Crossa *et al.* (1999) found that, in tropical maize, higher maximum temperature in low- and intermediate-altitude sites affected the expression of some QTL, whereas minimum temperature affected the expression of other QTL. Jiang *et al.* (1999) used molecular markers to investigate adaptation differences between highland and lowland tropical maize. They concluded that breeding for broad thermal adaptation should be possible by pooling genes showing adaptation to specific thermal regimes, albeit at the expense of reduced progress for specific adaptation. Molecular marker-assisted selection would be an ideal tool for this task because it could reduce linkage drag

caused by the unintentional transfer of undesirable traits (Jiang *et al.*, 1999).

Elias *et al.* (2016) have suggested that quantification of differential effects of segments of a genome across environments can be done by exploiting marker × environment (M×E) interactions. Crossa *et al.* (2016) suggested that the M×E genomic model can be used to generate predictions for untested individuals and identify genomic regions in which effects are stable across environments and others that show environmental specificity. They found that the M×E model performed better than the single-environment and across-environment models in minimizing the model residual variance. Jarquín *et al.* (2016) proposed a hierarchical Bayesian formulation of a linear–bilinear model, and showed that the proposed model facilitated identifying groups of genotypes and sites that caused GEI across years and within years, since the hierarchical Bayesian structure allowed using plant breeding data from different years by borrowing information among them.

Breeding for stability/reliability of performance

Evans (1993) pointed to the need for developing new cultivars with broad adaptation to a number of diverse environments (selection for adaptability) and to the need for farmers to use new cultivars with reliable or consistent performance from year to year (reliability). Studinicki *et al.* (2019) have recently pointed out that cultivar recommendation based on mean performance determined by METs conducted on research stations could be unreliable and ineffective for assessing performance in farmers' fields, and that it was important to improve the efficiency of cultivar recommendation based on METs.

Smith *et al.* (1990) indicated that genetic improvement for low-input conditions would require capitalizing on GEI and that slower or limited gains in low-input or stress environments suggested that conventional high-input management of breeding nurseries and evaluation trials might not effectively select genotypes with improved performance at low-input levels. This viewpoint was also highlighted by Ceccarelli *et al.* (2001). Because of the successes in favourable environments, plant breeders have tried to solve the problems of poor farmers living in unfavourable environments by simply extending the same methodologies and philosophies applied to favourable, high-potential environments, without considering the possible limitations associated with the presence of a large GEI (Ceccarelli *et al.*, 2001). Selection in good environments is favoured because it is believed that heritabilities are higher there than in poor environments (Blum, 1988). Singh and Ceccarelli (1995) suggested, however, that there was no relationship between yield level and magnitude of heritability. Rosielle and Hamblin (1981) examined theoretical aspects of selection for yield in stress and non-stress environments. They showed that selection for tolerance to stress generally reduced mean yield in non-stress environments and that selection for mean productivity generally increased mean yields in both stress and non-stress environments. Bramel-Cox (1996) reviewed relevant literature on breeding for reliability of performance in unpredictable environments.

To be reliable, a stability statistic must be based on a large number of environments (more than ten). Information on stability can usually be obtained in the final stages of a breeding programme, when replicated tests are conducted. From the standpoint of individual growers, stability across years (temporal) is most important. A breeder could test cultivars or lines for 10–15 years and identify those that have temporal stability. Crosses could then be made among the most stable cultivars to develop source material (germplasm) that would be utilized for developing inbred lines or pure lines. Therefore, extensive cultivar testing across years is a precursor to cultivar development.

Stability of cultivars would be enhanced if multiple resistances/tolerances to stress factors were incorporated into the germplasm used for cultivar development. If every cultivar (different genotypes) possessed equal resistance/tolerance to every major stress encountered in diverse target environments, GEI would be reduced. Conversely, if genotypes possessed differential levels of resistance (a heterogeneous group) and, somehow, we could make all target environments as homogeneous as possible, GEI would again be reduced. Since we do not have any control over unpredictable environments from year to year, the best approach would be the former.

Stability analyses can be used to identify durable resistance to disease pathogens (Jenns *et al.*, 1982). If a cultivar–pathogen-isolate interaction exists, it would be necessary to identify a cultivar that has general resistance instead of specific resistance. GGE biplot methodology has been used to study host genotype-by-pathogen strain interactions (Yan and Kang, 2003).

Kang *et al.* (1987a) examined whether stability of one trait was correlated with stability of another trait. If the stability (stability variance, ecovalence or any other stability statistic) of two traits was reasonably well, positively correlated, concurrent selection for stabilities of the two traits could be possible.

Measure interaction at intermediate growth stages

A crop is exposed to variable environmental factors throughout the growing season. Generally, researchers investigate the causes of GEI at the final harvest stage. To critically investigate and better understand GEI, one may need to record environmental variables and plant-growth measurements at weekly intervals. This would help determine what effect, if any, the environmental variables from an earlier period had on GEI at intermediary stages and on final yield. This may provide a better understanding of the dynamic process of yield formation.

Early multi-environment testing

Usually, there is a shortage of seed at the earliest stages of breeding, which prevents extensive testing. However, in a clonally propagated crop, such as sugarcane or potato, one stalk of sugarcane or one tuber of potato can be divided into at least two pieces and planted in more than one environment. Similarly, in other crops, if only 20 kernels are available, one could plant ten seeds each in two diverse environments. In the absence of a GEI, one would obtain a better evaluation of the genotypes, but, if GEI was present, one would obtain information about the consistency or inconsistency of performance of genotypes early in the programme. This strategy would prevent gene loss or genetic erosion, which could occur if testing was done in only one environment, and would also result in an increased breeding gain without a corresponding increase in expenditure of resources.

Optimal resource allocation

GEI can be employed to judiciously allocate resources in a breeding programme (Pandey and Gardner, 1992; Magari *et al.*, 1996). Carter *et al.* (1983) estimated that, at a low level of treatment–environment interaction (10% of error variance), testing in at least two environments was necessary to detect treatment differences of 20% and it required at least seven environments to detect smaller (10%) treatment differences for growth-analysis experiments in soybean. With a larger magnitude of interaction, a larger number of environments would be needed for a given level of precision in treatment differences.

Magari *et al.* (1996) used multi-environment (different planting dates) data for ear-moisture loss rate in maize, which exhibited planting date-by-genotype interaction. The relative efficiency for the benchmark protocol (11 plants per replication, three replications and three planting dates) was regarded as the reference value (100%). The relative efficiency for five plants per plot in four replications and three planting dates was equivalent to that for the benchmark protocol. A relative efficiency of 100% could also be achieved with a sample of four planting dates, three replications and three to four plants per plot. When the number of replications was increased to four in each of the four planting dates, only two plants per plot were needed to achieve a relative efficiency of 100%. The number of planting dates (environments) was found to be a critical factor in determining the precision of an experiment.

Outlook

To ensure the stability of crop production, the basic crop germplasm pools would need to be broadened (Sperling *et al.*, 2001). Duvick (2002) envisioned that the farmer-breeders (acting either as individuals or in associations, such as communities) and their non-governmental organization (NGO) partners would produce

varieties with utility in farming systems that are not well served (or not served at all) by formal plant breeding, either public or private. Thus, there should be a greater emphasis on participatory plant breeding, which involves scientists, farmers, consumers, extension personnel, industry and others, in the future. The article by Studinicki *et al.* (2019) is relevant here. Participatory plant breeding should be expanded, especially in developing countries. This should help broaden the genetic base of crops and stabilize food production as a result of farmers' developing, identifying and using locally adapted crop varieties that are farmer-acceptable and farmer-accessible. Decentralized or participatory plant breeding is essential for exploiting specific adaptation fully and making positive use of GEI, as pointed out by Ceccarelli *et al.* (2001).

In the previous edition of this book, climate change was rarely mentioned. Effects of climate change on crops have become more obvious now. Climate change, resulting from global warming caused by emissions of greenhouse gases, needs to be dealt with through mitigation and adaptation strategies. The role of plant breeders in identifying genes or QTL conditioning resistance/tolerance to biotic and abiotic stresses will be ever greater in the future. Molecular biology (including molecular genetics, biochemistry and plant physiology) will play an even greater role in breeding crop species and overcoming the constraints imposed on genotypes by their interaction with environmental factors. For example, cloning of genes for tolerance to cold, heat, salinity and other stresses, and their insertion into cultivars lacking in those genes, could overcome the mentioned stresses. Greater emphasis will need to be placed on identifying new sources of resistance/tolerance to tackle the extraordinary stresses imposed by climate change. Causes of GEI will need to be isolated on a case-by-case basis.

References

Aastveit, A.H. and Mejza, S. (1992) A selected bibliography on statistical methods for the analysis of genotype × environment interaction. *Biuletyn Oceny Odmian* 25, 83–97.

Annicchiarico, P. and Perenzin, M. (1994) Adaptation patterns and definition of macro-environments for selection and recommendation of common wheat genotypes in Italy. *Plant Breeding* 113, 197–205.

Bachmann, K., Chambers, K.L. and Price, H.J. (1985) Genome size and natural selection: Observations and experiments in plants. In: Cavalier-Smith, T. (ed.) *The Evolution of Genome Size*. Wiley, Chichester, UK, pp. 267–276.

Baker, R.J. (1988) Differential response to environmental stress. In: Weir, B.S., Eisen, E.J., Goodman, M.M. and Namkoong, G. (eds) *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer, Sunderland, Massachusetts, pp. 492–504.

Baker, R.J. (1990) Crossover genotype– environmental interaction in spring wheat. In: Kang, M.S. (ed.) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University Agricultural Center, Baton Rouge, Louisiana, pp. 42–51.

Basford, K.E. and Tukey, J.W. (1999) *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*. Chapman & Hall/ CRC, Boca Raton, Florida.

Beavis, W.D. and Keim, P. (1996) Identification of quantitative trait loci that are affected by environment. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 123–149.

Becker, H.C. (1981) Correlations among some statistical measures of phenotypic stability. *Euphytica* 30, 835–840.

Becker, H.C. and Leon, J. (1988) Stability analysis in plant breeding. *Plant Breeding* 101, 1–23.

Bennett, M.D. (1987) Variation in genomic form in plants and its ecological implications. *New Phytologist* 106(Suppl.), 177–200.

Blum, A. (1988) *Plant Breeding for Stress Environments*. CRC Press, Boca Raton, Florida.

Bramel-Cox, P.J. (1996) Breeding for reliability of performance across unpredictable environments. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 309–339.

Brancourt-Hulmel, M. (1999) Crop diagnosis and probe genotypes for interpreting genotype environment interaction in winter wheat trials. *Theoretical and Applied Genetics* 99, 1018–1030.

Brancourt-Hulmel, M., Denis, J.-B. and Lecomte, C. (2000) Determining environmental covariates which explain genotype environment interaction in winter wheat through probe genotypes and biadditive factorial regression. *Theoretical and Applied Genetics* 100, 285–298.

Carter, T.E., Jr, Burton, J.W., Cappy, J.J., Israel, D.W. and Boerma, H.R. (1983) Coefficients of variation, error variances, and resource allocation in soybean growth analysis experiments. *Agronomy Journal* 75, 691–696.

Cavalier-Smith, T. (1985a) Eucaryote gene number, non-coding DNA and genome size. In: Cavalier-Smith, T. (ed.) *The Evolution of Genome Size*. Wiley, Chichester, UK, pp. 69–103.

Cavalier-Smith, T. (1985b) Cell volume and the evolution of eukaryotic genome size. In: Cavalier-Smith, T. (ed.) *The Evolution of Genome Size*. Wiley, Chichester, UK, pp. 105–184.

Ceccarelli, M., Falistocco, E. and Cionini, P.G. (1992) Variation of genome size and organization within hexaploid *Festuca arundinacea*. *Theoretical and Applied Genetics* 83, 273–278.

Ceccarelli, S., Erskine, W., Hamblin, J. and Grando, S. (1994) Genotype by environment interaction and international breeding programmes. *Experimental Agriculture* 30, 177–187.

Ceccarelli, S., Grando, S., Amri, A., Asaad, F.A., Benbelkacem, A., *et al*. (2001) Decentralized and participatory plant breeding for marginal environments. In: Cooper, H.D., Spillane, C. and Hodgkins, T. (eds) *Broadening the Genetic Bases of Crop Production*. CAB International, Wallingford, UK, pp. 115–135.

Clark, R.B. and Duncan, R.R. (1993) Selection of plants to tolerate soil salinity, acidity, and mineral deficiencies. In: Bruxton, D.R., Shibles, R., Forsberg, R.A., Blad, B.A., Asay, K.H., *et al*. (eds) *International Crop Science* I. Crop Science Society of America, Madison, Wisconsin, pp. 371–379.

Cooper, M. and Hammer, G.L. (eds) (1996) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK, ICRISAT, Patancheru, India, and IRRI, Manila, The Philippines.

Cooper, M., Rajatasereekul, S., Immark, S., Fukai, S. and Basnayake, J. (1999) Rainfed lowland rice breeding strategies for Northeast Thailand. I. Genotypic variation and genotype × environment interactions for grain yield. *Field Crops Research* 64, 131–151.

Cooper, M., Technow, F., Messina, C., Gho, C. and Totir, L.R. (2016) Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Science* 56, 2141–2156. DOI: 10.2135/cropsci2015.08.0512.

Cornelius, P.L., Crossa, J. and Seyedsadr, M.S. (1996) Statistical tests and estimates of multiplicative models for GE interaction. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 199–234.

Crispeels, M.J. (ed.) (1994) *Introduction to Signal Transduction in Plants: A Collection of Updates*. American Society of Plant Physiologists, Rockville, Maryland.

Crossa, J., Cornelius, P.L. and Seyedsadr, M.S. (1996) Using the shifted multiplicative model cluster methods for crossover GE interaction. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 175–198.

Crossa, J., Vargas, M., van Eeuwijk, F.A., Jiang, C., Edmeades, G.O., *et al*. (1999) Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics* 99, 611–625.

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., *et al*. (2016) Extending the marker × environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science* 56, 2193–2209. DOI: 10.2135/cropsci2015.04.0260.

Crow, J.F. (1986) *Basic Concepts in Population, Quantitative and Evolutionary Genetics*. W.H. Freeman & Co., New York.

Dashiell, K.E., Ariyo, O.J. and Bello, L. (1994) Genotype × environment interaction and simultaneous selection for high yield and stability in soybeans (*Glycine max* (L.) Merr.). *Annals of Applied Biology* 124, 133–139.

DeLacy, I.H., Cooper, M. and Basford, K.E. (1996) Relationships among analytical methods used to study genotype-by-environment interactions and evaluation of their impact on response to selection. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 51–84.

de Leon, D., Jannink, J.-L., Edwards, J.W. and Kaeppler, S.M. (2016) Introduction to a special issue on genotype by environment interaction. *Crop Science* 56, 2081–2089. DOI: 10.2135/cropsci2016.07.0002in.

Denis, J.-B. (1988) Two-way analysis using covariates. *Statistics* 19, 123–132.

Denis, J.-B. and Gower, J.C. (1996) Asymptotic confidence regions for biadditive models: Interpreting genotype–environment interactions. *Applied Statistics* 45, 479–493.

Denis, J.-B., Gauch, H.G., Jr, Kang, M.S., Van Eeuwijk, F.A. and Zobel, R.W. (1996a) Bibliography on genotype-by-environment interaction. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 405–409.

Denis, J.-B., Piepho, H.-P. and van Eeuwijk, F.A. (1996b) *Mixed Models for Genotype by Environment Tables with an Emphasis on Heteroscedasticity*. Technical Report, Département de Biométrie, Laboratoire de Biométrie, INRA-Versailles, France, 23 pp.

Dickerson, G.E. (1962) Implications of genetic–environmental interaction in animal breeding. *Animal Production* 4, 47–64.

Doolittle, D.P. (1987) *Population Genetics: Basic Principles*. Springer, New York.

Duvick, D.N. (1992) Genetic contributions to advances in yield of US maize. *Maydica* 37, 69–79.

Duvick, D.N. (1996) Plant breeding, an evolutionary concept. *Crop Science* 36, 539–548.

Duvick, D.N. (2002) Crop breeding in the twenty-first century. In: Kang, M.S. (ed.) *Crop Improvement: Challenges in the Twenty-first Century*. Food Products Press, Binghamton, New York, pp. 1–15.

Eberhart, S.A. and Russell, W.A. (1966) Stability parameters for comparing varieties. *Crop Science* 6, 36–40.

Eisemann, R.L., Cooper, M. and Woodruff, D.R. (1990) Beyond the analytical methodology, better interpretation and exploitation of GE interaction in plant breeding. In: Kang, M.S. (ed.) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University Agricultural Center, Baton Rouge, Louisiana, pp. 108–117.

Elias, A.A., Robbins, K.R., Doerge, R.W. and Tuinstra, M.R. (2016) Half a century of studying genotype × environment interactions in plant breeding experiments. *Crop Science* 56, 2090–2105. DOI: 10.2135/cropsci2015.01.0061

Eskridge, K.M. (1996) Analysis of multiple environment trials using the probability of outperforming a check. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 273–307.

Evans, L.T. (1993) *Crop Evolution, Adaptation and Yield*. Cambridge University Press, New York.

Falconer, D.S. (1952) Selection for large and small size in mice. *Journal of Genetics* 51, 470–501.

Federer, W.T. and Scully, B.T. (1993) A parsimonious statistical design and breeding procedure for evaluating and selecting desirable characteristics over environments. *Theoretical and Applied Genetics* 86, 612–620.

Fernandez, G.C.J. (1991) Analysis of genotype × environment interaction by stability estimates. *HortScience* 26, 947–950.

Finlay, K.W. and Wilkinson, G.N. (1963) The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research* 14, 742–754.

Flores, F., Moreno, M.T. and Cubero, J.I. (1998) A comparison of univariate and multivariate methods to analyze G × E interaction. *Field Crops Research* 56, 271–286.

Freeman, G.H. (1975) Analysis of interactions in incomplete two-way tables. *Applied Statistics* 24, 46–55.

Freeman, G.H. and Perkins, J.M. (1971) Environmental and genotype–environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. *Heredity* 27, 15–23.

Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.

Gauch, H.G., Jr (1992) *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam, The Netherlands.

Gauch, H.G., Jr and Furnas, R.E. (1991) Statistical analysis of yield trials with MATMODEL. *Agronomy Journal* 83, 916–920.

Gauch, H.G., Jr and Moran, D.R. (2019) AMMISOFT for AMMI analysis with best practices. bioRxiv preprint http://dx.doi.org/10.1101/538454.

Gauch, H.G., Jr and Zobel, R.W. (1990) Imputing missing yield trial data. *Theoretical and Applied Genetics* 70, 753–761.

Gauch, H.G., Jr and Zobel, R.W. (1996) AMMI analysis of yield trials. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 85–122.

Glaz, B. and Dean, J.L. (1988) Statistical error rates and their implications in sugarcane clone trials. *Agronomy Journal* 80, 560–562.

Glaz, B. and Kang, M.S. (2008). Location contributions determined via GGE biplot analysis of multienvironment sugarcane genotype-performance trials. *Crop Science* 48, 941–950.

Glaz, B., Miller, J.D. and Kang, M.S. (1985) Evaluation of cultivar-testing locations in sugarcane. *Theoretical and Applied Genetics* 71, 22–25.

Gorman, D.P., Kang, M.S. and Milam, M.R. (1989) Contribution of weather variables to genotype × environment interaction in grain sorghum. *Plant Breeding* 103, 299–303.
Gregorius, H.R. and Namkoong, G. (1986) Joint analysis of genotypic and environmental effects. *Theoretical and Applied Genetics* 72, 413–422.
Grogan, S.M., Anderson, J., Baenziger, P.S., Frels, K., Guttieri, M.J., *et al*. (2016) Phenotypic plasticity of winter wheat heading date and grain yield across the US Great Plains. *Crop Science* 56, 2223–2236. DOI: 10.2135/cropsci2015.06.0357.
Haji, H.M. and Hunt, L.A. (1999) Genotype × environment interactions and underlying environmental factors for winter wheat in Ontario. *Canadian Journal of Plant Science* 79, 497–505.
Haldane, J.B.S. (1947) The interaction of nature and nurture. *Annals of Eugenics* 13, 197–205.
Hall, A.E. (2001) *Crop Responses to Environment*. CRC Press, Boca Raton, Florida.
Hardwick, R.C. and Wood, J.T. (1972) Regression methods for studying genotype–environment interactions. *Heredity* 28, 209–222.
Hayes, B.J., Daetwyler, H.D. and Goddard, M.E. (2016) Models for genome × environment interaction: Examples in livestock. *Crop Science* 56, 2251–2259. DOI: 10.2135/cropsci2015.07.0451.
Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447.
Herrera-Estrella, L. and Simpson, J. (1990) Influence of environmental factors on photosynthetic genes. In: Scandalios, J.G. and Wright, T.R.F. (eds) *Advances in Genetics*. Academic Press, New York, pp. 133–163.
Higley, L.G., Browde, J.A. and Higley, P.M. (1993) Moving toward new understandings of biotic stress and stress interactions. In: Bruxton, D.R., Shibles, R., Forsberg, R.A., Blad, B.L., Asay, K.H., *et al*. (eds) *International Crop Science* I. Crop Science Society of America, Madison, Wisconsin, pp. 749–754.
Hildebrand, P.E. and Russell, J.T. (1996) *Adaptability Analysis: A Method for the Design, Analysis and Interpretation of On-farm Research-extension*. Iowa State University Press, Ames, Iowa.
Hoffmann, A.A. and Parsons, P.A. (1997) *Extreme Environmental Change and Evolution*. Cambridge University Press, Cambridge, UK.
Hühn, M. (1996) Nonparametric analysis of genotype × environment interactions by ranks. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 235–271.
Hussein, M.A., Bjornstad, A. and Aastveit, A.H. (2000) SASG × ESTAB: A SAS program for computing genotype × environment stability statistics. *Agronomy Journal* 92, 454–459.
Jarquín, D., Pérez-Elizalde, S., Burgueño, J. and Crossa, J. (2016) A hierarchical Bayesian estimation model for multienvironment plant breeding trials in successive years. *Crop Science* 56, 2260–2276. DOI: 10.2135/cropsci2015.08.0475.
Jenns, A.E., Leonard, K.J. and Moll, R.H. (1982) Stability analyses for estimating relative durability of quantitative resistance. *Theoretical and Applied Genetics* 63, 183–192.
Jiang, C., Edmeades, G.O., Armstead, I., Lafitte, H.R., Hayward, M.D. and Hoisington, D. (1999) Genetic analysis of adaptation differences between highland and lowland tropical maize using molecular markers. *Theoretical and Applied Genetics* 99, 1106–1119.
Johnson, J.J., Alldredge, J.R., Ullrich, S.E. and Dangi, O. (1992) Replacement of replications with additional locations for grain sorghum cultivar evaluation. *Crop Science* 32, 43–46.
Kang, M.S. (1988) A rank-sum method for selecting high-yielding, stable corn genotypes. *Cereal Research Communications* 16, 113–115.
Kang, M.S. (ed.) (1990) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University Agricultural Center, Baton Rouge, Louisiana.
Kang, M.S. (1993a) Issues in GE interaction. In: Rao, V., Hanson, I.E. and Rajanaidu, N. (eds) *Genotype–Environment Interaction Studies in Perennial Tree Crops*. Palm Oil Research Institute of Malaysia, Kuala Lumpur, pp. 67–73.
Kang, M.S. (1993b) Simultaneous selection for yield and stability in crop performance trials: Consequences for growers. *Agronomy Journal* 85, 754–757.
Kang, M.S. (1998) Using genotype-by-environment interaction for crop cultivar development. *Advances in Agronomy* 62, 199–252.
Kang, M.S. and Gauch, H.G., Jr (eds) (1996) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida.
Kang, M.S. and Gorman, D.P. (1989) Genotype × environment interaction in maize. *Agronomy Journal* 81, 662–664.

Kang, M.S. and Magari, R. (1995) STABLE: Basic program for calculating yield–stability statistic. *Agronomy Journal* 87, 276–277.
Kang, M.S. and Magari, R. (1996) New developments in selecting for phenotypic stability in crop breeding. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 1–14.
Kang, M.S. and Miller, J.D. (1984) Genotype × environment interactions for cane and sugar yield and their implications in sugarcane breeding. *Crop Science* 24, 435–440.
Kang, M.S. and Pham, H.N. (1991) Simultaneous selection for high yielding and stable crop genotypes. *Agronomy Journal* 83, 161–165.
Kang, M.S., Glaz, B. and Miller, J.D. (1987a) Interrelationships among stabilities of important agronomic traits in sugarcane. *Theoretical and Applied Genetics* 74, 310–316.
Kang, M.S., Miller, J.D. and Darrah, L.L. (1987b) A note on relationship between stability variance and ecovalence. *Journal of Heredity* 78, 107.
Kang, M.S., Harville, B.G. and Gorman, D.P. (1989) Contribution of weather variables to genotype × environment interaction in soybean. *Field Crops Research* 21, 297–300.
Kang, M.S., Gorman, D.P. and Pham, H.N. (1991) Application of a stability statistic to international maize yield trials. *Theoretical and Applied Genetics* 81, 162–165.
Khush, G.S. (1993) Breeding rice for sustainable agricultural systems. In: Buxton, D.R., Shibles, R., Forsberg, R.A., Blad, B.L., Asay, K.H., *et al*. (eds) *International Crop Science* I. Crop Science Society of America, Madison, Wisconsin, pp. 189–199.
Kleinknecht, K., Möhring, J., Laidig, F., Meyerb U. and Piepho, H.P. (2016) A simulation-based approach for evaluating the efficiency of multienvironment trial designs. *Crop Science* 56, 2237–2250. DOI: 10.2135/cropsci2015.07.0405.
Lado, B., Barrios, P.G., Quincke, M., Silva, P. and Gutiérrez, L. (2016) Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Science* 56, 2165–2179. DOI: 10.2135/cropsci2015.04.0207.
Laurie, D.A. and Bennett, M.D. (1985) Nuclear DNA content in the genera *Zea* and *Sorghum*: Intergeneric, interspecific and intraspecific variation. *Heredity* 55, 307–313.
Lee, E.A., Deen, W., Hooyer, M.E., Chambers, A., Parkin, G., *et al*. (2016) Involvement of year-to-year variation in thermal time, solar radiation and soil available moisture in genotype-by-environment effects in maize. *Crop Science* 56, 2180–2192. DOI: 10.2135/cropsci2015.04.0231.
Lee, M. (1995) DNA markers and plant breeding programs. *Advances in Agronomy* 55, 265–344.
Lin, C.S. and Binns, M.R. (1988) A method of analyzing cultivar × location × year experiments: A new stability parameter. *Theoretical and Applied Genetics* 76, 425–430.
Lin, C.S., Binns, M.R. and Lefkovitch, L.P. (1986) Stability analysis: Where do we stand? *Crop Science* 26, 894–900.
Lin, C.Y. and Lin, C.S. (1994) Investigation of genotype–environment interaction by cluster analysis in animal experiments. *Canadian Journal of Animal Science* 74, 607–612.
McKeand, S.E., Li, B., Hatcher, A.V. and Weir, R.J. (1990) Stability parameter estimates for stem volume for loblolly pine families growing in different regions in the southeastern United States. *Forest Science* 36, 10–17.
Magari, R., Kang, M.S. and Zhang, Y. (1996) Sample size for evaluating field ear moisture loss rate in maize. *Maydica* 41, 19–24.
Magari, R., Kang, M.S. and Zhang, Y. (1997) Genotype by environment interaction for ear moisture loss rate in corn. *Crop Science* 37, 774–779.
Malosetti, M., Bustos-Korts, D., Boer, M.P. and van Eeuwijk, F.A. (2016) Predicting responses in multiple environments: Issues in relation to genotype × environment interactions. *Crop Science* 56, 2210–2222. DOI: 10.2135/cropsci2015.05.0311.
Montaldo, H.H. (2001) Genotype by environment interactions in livestock breeding programs: A review. *Interciencia* 26(6), 229–235.
Mooers, C.A. (1921) The agronomic placement of varieties. *Journal of American Society of Agronomy* 13, 337–352.
Ougham, H.J. and Howarth, C.J. (1988) Temperature shock proteins in plants. In: Long, S.P. and Woodward, F.J. (eds) *Plants and Temperature Symposium of the Society for Experimental Biology*. Company of Biologists, Cambridge, UK, pp. 259–280.
Paderewski, J., Gauch, H.G. Jr, Mądry, W. and Gacek, E. (2016) AMMI analysis of four-way genotype × location × management × year data from a wheat trial in Poland. *Crop Science* 56, 2157–2164. DOI: 10.2135/cropsci2015.03.0152.

Pandey, S. and Gardner, C.O. (1992) Recurrent selection for population, variety, and hybrid improvement in tropical maize. *Advances in Agronomy* 48, 1–87.

Pauli, D., Chapman, S.C., Bart, R., Topp, C.N., Lawrence-Dill, C.J., Poland, J., *et al*. (2016) The quest for understanding phenotypic variation via integrated approaches in the field environment. *Plant Physiology* 172, 622–634. DOI: 10.1104/pp.16.00592.

Pazdernik, D.L., Hardman, L.L. and Orf, J.H. (1997) Agronomic performance and stability of soybean varieties grown in three maturity zones of Minnesota. *Journal of Production Agriculture* 10, 425–430.

Pedersen, A.R., Everson, E.H. and Grafius, J.E. (1978) The gene pool concept as basis for cultivar selection and recommendation. *Crop Science* 18, 883–886.

Piepho, H.-P. (1994) Missing observations in analysis of stability *Heredity* 72, 141–145. (Correction 73 [1994], 58.)

Piepho, H.-P. (1998) Methods of comparing the yield stability of cropping systems – a review. *Journal of Agronomy and Crop Science* 180, 193–213.

Piepho, H.-P. (2000a) Exact confidence limits for covariate-dependent risk in cultivar trials. *Journal of Agricultural Biological Environmental Statistics* 5, 202–213.

Piepho, H.-P. (2000b) A mixed model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* 156, 2043–2050.

Piepho, H.-P., Denis, J.B. and van Eeuwijk, F.A. (1998) Predicting cultivar differences using covariates. *Journal of Agricultural Biological Environmental Statistics* 3, 151–162.

Piepho, H.-P., Nazir, M.F., Qamar, M., Rattu, A., Riaz-ud-Din, *et al*. (2016) Stability analysis for a countrywide series of wheat trials in Pakistan. *Crop Science* 56, 2465–2475. DOI: 10.2135/cropsci2015.12.0743.

Pollack, C.J., Eagles, C.F., Howarth, C.J., Schunmann, P.H.D. and Stoddart, J.L. (1993) Temperature stress. In: Fowden, L., Mansfield, T. and Stoddart, J. (eds) *Plant Adaptation to Environmental Stress*. Chapman & Hall, New York, pp. 109–132.

Polle, A. and Rennenberg, H. (1993) Significance of antioxidants in plant adaptation to environmental stress. In: Fowden, L., Mansfield, T. and Stoddart, J. (eds) *Plant Adaptation to Environmental Stress*. Chapman & Hall, New York, pp. 263–273.

Prabhakaran, V.T. and Jain, J.P. (1994) *Statistical Techniques for Studying Genotype– Environment Interactions*. South Asian Publishers, New Delhi, India.

Rameau, C. and Denis, J.-B. (1992) Characterization of environments in long-term multi-site trials in asparagus, through yield of standard varieties and use of environmental covariates. *Plant Breeding* 109, 183–191.

Rao, T.D.P., Rao, D.V.S. and Rai, S.C. (1988) Symposium on statistical aspects of stability of crop yields. *Journal of Indian Society of Agricultural Statistics* 60, 70–79.

Rao, V., Henson, I.E. and Rajanaidu, N. (eds) (1993) *Genotype × Environment Interaction in Perennial Tree Crops*. International Society of Oil Palm Breeders and Palm Oil Research Institute of Malaysia, Kuala Lumpur.

Rayburn, A.L. and Auger, J.A. (1990) Genome size variation in *Zea mays* ssp. *mays* adapted to different altitudes. *Theoretical and Applied Genetics* 79, 470–474.

Rayburn, A.L., Price, H.J., Smith, J.D. and Gold, J.R. (1985) C-banded heterochromatin and DNA content in *Zea mays*. *American Journal of Botany* 72, 1610–1617.

Reckling, M., Döring, T.F., Stein-Bachinger, K., Bloch, R. and Bachinger, J. (2015) Yield stability of grain legumes in an organically managed monitoring experiment. *Aspects of Applied Biology* 128, 57–62.

Redei, G.P. (1982) *Genetics*. Macmillan, New York.

Rodrigues, P.C. (2018) An overview of statistical methods to detect and understand genotype-by-environment interaction and QTL-by-environment interaction. *Biometrical Letters* 55(2), 123–138. DOI: 10.2478/bile-2018-0009.

Rosielle, A.A. and Hamblin, J. (1981) Theoretical aspects of selection for yield in stress and non-stress environments. *Crop Science* 21, 943–946.

Ryals, J., Uknes, S. and Ward, E. (1994) Systemic acquired resistance. *Plant Physiology* 104, 1109–1112.

Saeed, M. and Francis, C.A. (1984) Association of weather variables with genotype × environment interaction in grain sorghum. *Crop Science* 24, 13–16.

Scandalios, J.G. (1990) Response of plant antioxidant defense genes to environmental stress. In: Scandalios, J.G. and Wright, T.R.F. (eds) *Advances in Genetics*. Academic Press, New York, pp. 1–41.

Schlichting, C.D. (1986) The evolution of phenotypic plasticity in plants. *Annual Reviews of Ecological Systematics* 17, 667–693.

Searle, S.R. (1987) *Linear Models for Unbalanced Data*. Wiley, New York.

Shukla, G.K. (1972) Some statistical aspects of partitioning genotype–environmental components of variability. *Heredity* 29, 237–245.

Silvey, V. (1981) The contribution of new wheat, barley and oat varieties to increasing yield in England and Wales 1947–78. *Journal of National Institute of Agricultural Botany* 15, 399–412.

Simmonds, N.W. (1981) Genotype (G), environment (E) and GE components of crop yields. *Experimental Agriculture* 17, 355–362.

Singh, M. and Ceccarelli, S. (1995) Estimation of heritability using varietal trials data from incomplete blocks. *Theoretical and Applied Genetics* 90, 142–145.

Singh, M., Ceccarelli, S. and Grando, S. (1999) Genotype × environment interaction of crossover type: Detecting its presence and estimating the crossover point. *Theoretical and Applied Genetics* 99, 988–995.

Smith, M.E., Coffman, W.R. and Barker, T.C. (1990) Environmental effects on selection under high and low input conditions. In: Kang, M.S. (ed.) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University Agricultural Center, Baton Rouge, Louisiana, pp. 261–272.

Specht, J.E. and Laing, D.R. (1993) Selection for tolerance to abiotic stresses – discussion. In: Bruxton, D.R., Shibles, R., Forsberg, R.A., Blad, B.L., Asay, K.H., *et al*. (eds) *International Crop Science* I. Crop Science Society of America, Madison, Wisconsin, pp. 381–382.

Sperling, L., Ashby, J., Weltzien, E., Smith, M. and McGuire, S. (2001) Base-broadening for client-oriented impact: insights drawn from participatory plant breeding field experience. In: Cooper, H.D., Spillane, C. and Hodgkins, T. (eds) *Broadening the Genetic Bases of Crop Production*. CAB International, Wallingford, UK, pp. 419–435.

Studinicki, M., Kang, M.S., Iwańska, M., Oleksiak, T., Wójcik-Grant, E. *et al*. (2019) Consistency of yield ranking and adaptability patterns of winter wheat cultivars between multi-environmental trials and farmer surveys. *Agronomy* 9, 245–254. DOI: 10.3390/agronomy9050245.

Tai, G.C.C. (1971) Genotypic stability analysis and its application to potato regional trials. *Crop Science* 11, 184–190.

Tai, G.C.C. (1990) Path analysis of genotype-environment interactions. In: Kang, M.S. (ed.) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University Agricultural Center, Baton Rouge, Louisiana, pp. 273–286.

Tai, G.C.C. and Coleman, W.K. (1999) Genotype × environment interaction of potato chip colour. *Canadian Journal of Plant Science* 79, 433–438.

Unsworth, M.H. and Fuhrer, J. (1993) Crop tolerance to atmospheric pollutants. In: Bruxton, D.R., Shibles, R., Forsberg, R.A., Blad, B.L., Asay, K.H., *et al*. (eds) *International Crop Science* I. Crop Science Society of America, Madison, Wisconsin, pp. 363–370.

van Eeuwijk, F.A., Denis, J.-B. and Kang, M.S. (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 15–49.

van Eeuwijk, F.A., Bustos-Korts, D.V. and Malosetti, M. (2016) What should students in plant breeding know about the statistical aspects of genotype × environment interactions? Crop Science 56, 2119–2140. DOI: 10.2135/cropsci2015.06.0375.

Vargas, M., Crossa, J., van Eeuwijk, F.A., Sayre, K.D. and Reynolds, M.P. (2001) Interpreting treatment × environment interaction in agronomy trials. *Agronomy Journal* 93, 949–960.

Via, S. (1984) The quantitative genetics of polyphagy in an insect herbivore. I. Genotype–environment interaction in larval performance of different host plant species. *Evolution* 38, 881–895.

Wade, L.J., McLaren, C.G., Quintana, L., Harnpichitvitaya, D., Rajatasereekul, S., *et al*. (1999) Genotype by environment interactions across diverse rainfed lowland rice environments. *Field Crops Research* 64, 35–50.

Wang, D.L., Zhu, J., Li, Z.K. and Paterson, A.H. (1999) Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theoretical and Applied Genetics* 99, 1255–1264.

Weber, W.E., Wricke, G. and Westermann, T. (1996) Selection of genotypes and prediction of performance by analyzing GE interactions. In: Kang, M.S. and Gauch, H.G., Jr (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 353–371.

Witcombe, J.R. (2001) The impact of decentralized and participatory plant breeding on the genetic base of crops. In: Cooper, H.D., Spillane, C. and Hodgkins, T. (eds) *Broadening the Genetic Bases of Crop Production*. CAB International, Wallingford, UK, pp. 407–417.

Wood, J.T. (1976) The use of environmental variables in the interpretation of genotype–environment interaction. *Heredity* 37, 1–7.

Woodend, J.J. and Glass, A.D.M. (1993) Genotype–environment interaction and correlation between vegetative and grain production measures of potassium use-efficiency in wheat (*T. aestivum* L.) grown under potassium stress. *Plant and Soil* 151, 39–44.

Wricke, G. (1962) Über eine Methode zur Erfassung der ökologischen Streubreite in Feldversuchen. *Zeitschrift für Pflanzenzüchtung* 47, 92–96.

Xu, Y. (2016) Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics* 129(4), 653–673. DOI: 10.1007/s00122-016-2691-5.

Yan, W. (2001) GGEbiplot – a Windows application for graphical analysis of multi-environment trial data and other types of two-way data. *Agronomy Journal* 93(5), 1111–1118.

Yan, W. (2014) *Crop Variety Trials: Data Management and Analysis*. Wiley, Chichester, UK.

Yan, W. (2016) Analysis and handling of G × E in a practical breeding program. *Crop Science* 56, 2106–2118. DOI: 10.2135/cropsci2015.06.0336.

Yan, W. and Hunt, L.A. (2001) Genetic and environmental causes of genotype by environment interaction for winter wheat yield in Ontario. *Crop Science* 41, 19–25.

Yan, W. and Hunt, L.A. (2002) Biplot analysis of diallel data. *Crop Science* 42, 21–30.

Yan, W. and Rajcan, I. (2002) Biplot analysis of test sites and trait relations of soybean in Ontario. *Crop Science* 42, 11–20.

Yan, W., Hunt, L.A., Sheng, Q. and Szlavnics, Z. (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science* 40(3), 597–605.

Yan, W., Cornelius, P.L., Crossa, J. and Hunt, L.A. (2001) Two types of GGE biplots for analyzing multi-environment trial data. *Crop Science* 41(3), 656–663.

Yan, W. and Kang, M.S. (2003) *GGE biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC Press, Boca Raton, Florida.

Yates, F. and Cochran, W.G. (1938) The analysis of groups of experiments. *Journal of Agricultural Science* 28, 556–580.

Zavala-Garcia, F. and Treviño-Hernández, T.E. (eds) (2000) *Simposium interaccion genotipo × ambiente*. SOMEFI-CSSA-UG, Irapuato, Gto, Mexico.

Zhang, Q. and Geng, S. (1986) A method of estimating varietal stability for long-term trials. *Theoretical and Applied Genetics* 71, 810–814.

10 Biplot Analysis of Multi-environment Trial Data

Weikai Yan[1]* and L.A. Hunt[2]
[1]960 Carling Ave, Ottawa, Ontario, Canada; [2]University of Guelph, Guelph, Ontario, Canada

Introduction

Regional multi-environment trials (METs) are conducted every year for all major crops throughout the world, constituting a costly but essential step towards new crop cultivar release and recommendation. METs are essential because the presence of genotype–environment interaction (GEI), i.e. differential genotype responses in different environments, complicates cultivar evaluation. Some important concepts, such as ecological region, ecotype, mega-environment, specific adaptation and stability, all originate from GEI. As stated by Gauch and Zobel (1996), were there no GEI, a single cultivar would prevail all over the world and a single trial would suffice for cultivar evaluation. GEI constitutes a major challenge to cultivar improvement, and MET data analysis constitutes an important aspect of plant breeding. Because of this, improvement in the methods used for MET data analysis is of great interest to the plant-breeding community. This chapter deals with the biplot method, which has received much attention in recent years.

Multi-environment trial data analysis

The primary objective of an MET is, of course, to identify superior cultivars. The most common practice used to achieve this end is to compare the mean yield of genotypes across test environments (usually year–location combinations) represented in the MET. The validity of this practice is, however, based on the usually unstated assumption that the environments in the MET belong to a single mega-environment, defined as a group of locations in which the same set of cultivars perform best across a number of years. Cultivar evaluation is always specific to a single mega-environment. If the test environments are sufficiently heterogeneous, the cultivars that are selected based on mean yield may not be the best in some of the test environments; in extreme cases where GEI is dominant, they may even not be the best in any of the environments. Thus, a second utility of MET data analysis should be to investigate the relationships among the test environments and, if possible, to divide the target region into meaningful mega-environments. Identification of mega-environments would allow exploitation of the GEI that is repeatable across years.

For a given mega-environment, genotypes should be evaluated for mean yield (or, in more general terms, mean performance) and stability across test environments. The ideal cultivar should be one that is both high-yielding and stable. Mean performance is simply the mean across all environments, whereas stability is a

* Email: weikai.yan@canada.ca

© CAB International, 2020.

measure of variability across environments. Much research has focused on quantification of stability, and numerous stability measures have been proposed (Lin *et al.*, 1986; Lin and Binns, 1994; Kang, 1998). For a given mega-environment and parallel to cultivar evaluation, individual test environments should be evaluated for their ability to provide data that allow for discrimination among genotypes and, at the same time, for the extent to which they represent the target mega-environment.

The ultimate reason for differential stability among genotypes and for differential results from various test environments is non-repeatable GEI. Because this type of GEI cannot be effectively exploited, it must be avoided. A fourth utility of MET data analysis is the development of a better understanding of the causes of GEI. Such an understanding may help to avoid confounding plant responses to specific and rare conditions with overall cultivar evaluation.

To summarize, MET data analysis should, and potentially can, fulfil four functions: (i) investigation of possible mega-environment differentiation in the target environment; (ii) selection of superior cultivars for individual mega-environments; (iii) selection of better test environments; and (iv) development of a better understanding of the causes of GEI. An ideal MET data-analysis system should accomplish all four tasks so that the information contained in the MET is maximally exploited and utilized.

Visualization of multi-environment trial data

With the belief that 'a picture is worth a thousand words', many attempts have been made to graphically present MET data. The general pattern of such a graphical display of MET data is to plot the mean yield of each genotype against a measure of stability, which can be any parameter that is listed in Lin *et al.* (1986), among others.

Another popular presentation of MET data is based on the Finlay and Wilkinson (1963) model, in which the yield of each genotype is plotted against the mean yield of each environment and in which each genotype is represented by a fitted straight line. Philosophically, this type of graphical display of MET data is attractive, since it clearly indicates differential genotype responses to test environments. The problem with this method is that the environmental means are not always a good, and are frequently a poor, measure of environments, such that the fitted lines in most cases only account for a small fraction of the total GEI (Zobel *et al.*, 1988).

A visualization method that is similar to that of Finlay and Wilkinson (1963) but explains more GEI was developed by Gauch and Zobel (1997). In this method, the nominal yields of genotypes are plotted against the first interaction principal component (IPC1) scores of environments, so that each genotype is represented by a line with the mean yield as the intercept and the genotype IPC1 score as the slope. Such a plot indicates the 'which-won-where' patterns of the data, provided that the IPC1 explains most of the GEI.

The recently developed GGE-biplot method (Yan *et al.*, 2000, 2001) provides an elegant and highly useful display of MET data. It effectively addresses both the issue of mega-environment differentiation and the issue of genotype selection for a given mega-environment based on mean yield and stability. It also allows environments to be evaluated just as well as genotypes. In addition, it facilitates interpretation of GEI as a genotypic factor by environmental factor interaction (Yan and Hunt, 2001). In the rest of the chapter, we shall describe the rationale and applications of the GGE-biplot methodology in MET data analysis.

The GGE-biplot Methodology

The concept of biplot

The concept of biplot was first proposed by Gabriel (1971). The main ideas follow. Any two-way table or matrix X that contains n rows and m columns can be regarded as the product of two matrices: matrix A with n rows and r columns, and matrix B with r rows and m columns. Therefore, matrix X can always be decomposed into its two component matrices A and B. If r happens to be 2, matrix X is referred to as a rank-two matrix. Each row in matrix A has two values that can be displayed as a point in a two-dimensional plot. Similarly, each column in matrix B has two values and can also be displayed as a point in a two-dimensional plot. When both the n rows of

A and the *m* columns of B are displayed in a single plot, the plot is called a 'biplot'. Therefore, the biplot of a rank-two matrix contains *n* + *m* points, as compared with *n* × *m* values in the matrix per se, and yet contains all the information of the matrix.

One interesting property of a biplot is that each of the *n* × *m* values can be precisely recovered by viewing the *n* + *m* points on the biplot. Assume that we have yield data for three genotypes and three environments and that it is a rank-two matrix. After decomposition of the data into its two component matrices, the three genotypes and three environments can be presented in a biplot, as shown in Fig. 10.1.

The yield of genotype *i* in environment *j*, Y_{ij}, can be recovered by the following formula:

$$Y_{ij} = \overline{OE_j}\cos(\alpha_{ij})\overline{OG_i} = \overline{OE_j}\,\overline{OP_{ij}}$$

where $\overline{OG_i}$ is the absolute distance from the biplot origin *O* to the marker of the genotype *i*, $\overline{OE_j}$ is the absolute distance from the biplot origin *O* to the marker of environment *j*, α_{ij} is the angle between the vectors $\overline{OG_i}$ and $\overline{OE_j}$, and $\overline{OP_{ij}} = \cos\alpha_{ij}\overline{OG_i}$, which is the projection of the marker of genotype *i* on to the vector of environment *j*. To compare yields of the three genotypes in environment E_1, we have:

$$Y_{11} = \overline{OE_1}\cos(\alpha_{11})\overline{OG_1} = \overline{OE_1}\,\overline{OP_{11}}$$

$$Y_{21} = \overline{OE_1}\cos(\alpha_{21})\overline{OG_2} = \overline{OE_1}\,\overline{OP_{21}}$$

$$Y_{31} = \overline{OE_1}\cos(\alpha_{31})\overline{OG_3} = \overline{OE_1}\,\overline{OP_{31}}$$

where OP_{11}, OP_{21} and OP_{31} are the projections of the markers of the genotypes on to the vector or its extension of environment E_1. Because OE_1 is non-negative and common to all genotypes, comparisons among Y_{11}, Y_{21} and Y_{31} can be performed by simply visualizing OP_{11}, OP_{21} and OP_{31}. In our example (Fig. 10.1), it is obvious that $OP_{11} > OP21 > OP_{31}$, and therefore, $Y_{11} > Y_{21} > Y_{31}$. Note that OP_{11} and OP_{21} are above average, whereas OP_{31} is below average, since $cos(\alpha_{11})$ and $cos(\alpha_{21})$ are positive, and $cos(\alpha_{31})$ is negative.

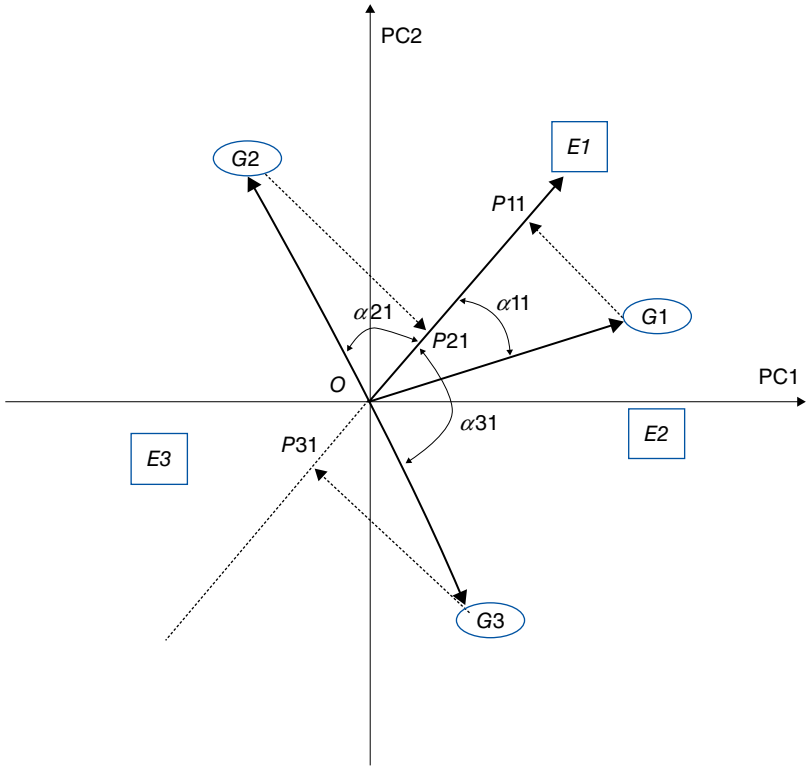


Fig. 10.1. The geometry of the biplot.

Approximation of any two-way table using a rank-two matrix

A biplot is obviously an elegant and useful display of a rank-two matrix. In reality, however, it is rare that a two-way dataset is exactly a rank-two matrix. Nevertheless, if a two-way dataset, e.g. the yield data of a number of genotypes tested in a number of environments, can be approximated by a rank-two matrix, the latter can then be displayed in a biplot (Gabriel, 1971). The process of decomposing matrix X into its component matrices A and B is called 'singular value decomposition' (SVD), the result of which is r principal components (r equals the smaller of n and m). If the first two principal components (PC1 and PC2) explain a large proportion of the total variation of X, X is said to be sufficiently approximated by a rank-two matrix and can be approximately displayed in a biplot.

The concept of GGE

The concept of GGE originates from analysis of METs of crop cultivars. The yield of a genotype (or any other measure of genotype performance) in an environment is a joint effect of genotype main effect (G), environment main effect (E) and GEI. In a normal MET, E accounts for 80% and G and GEI each account for about 10% of the total variation. For the purpose of cultivar evaluation, however, only G and GEI are pertinent (Gauch and Zobel, 1996). Furthermore, both G and GEI must be considered in cultivar evaluation; hence the term 'GGE' (Yan *et al.*, 2000). Simultaneous examination of G and GEI is, thus, an important principle in cultivar evaluation (Kang, 1993).

Models for constructing a GGE biplot

The GGE biplot displays the GGE part of an MET dataset. Compared with other types of biplots, a GGE biplot has these advantages: (i) it displays most information that is relevant to cultivar evaluation; and (ii) it displays only the information that is relevant to cultivar evaluation.

A GGE biplot can be generated based on SVD of: (i) environment-centred data; (ii) environment-centred and within-environment standard deviation-scaled data; and (iii) environment-centred and within-environment standard error-scaled data.

Singular value decomposition of environment-centred data

The model for a GGE biplot based on SVD of environment-centred data is:

$$Y_{ij} - \overline{Y_j} = \lambda_1 \xi_{i1} \eta_{j1} + \lambda_2 \xi_{i2} \eta_{j2} + \varepsilon_{ij} \quad \text{(Eqn 10.1a)}$$

where:
Y_{ij} is the mean yield of genotype i in environment j;
$\overline{Y_j}$ is the mean yield across all genotypes in environment j;
λ_1 and λ_2 are the singular values for the first and second principal components, PC1 and PC2, respectively;
ξ_{i1} and ξ_{i2} are the PC1 and PC2 scores, respectively, for genotype i;
η_{j1} and η_{j2} are the PC1 and PC2 scores, respectively, for environment j;
ε_{ij} is the residual of the model associated with genotype i in environment j.

To display the PC1 and PC2 in a biplot, the equation is rewritten as:

$$Y_{ij} - \overline{Y_j} = (\lambda_1^f \xi_{i1})(\lambda_1^{1-f} \eta_{j1}) + (\lambda_2^f \xi_{i2})(\lambda_2^{1-f} \eta_{j2}) + \varepsilon_{ij} \quad \text{(Eqn 10.1b)}$$

where $f = 1$ or 0, which is the singular value partitioning factor. When $f = 1$, noted as SVP = 1, it is referred to as genotype-focused partitioning, and the biplot is most suitable for comparing genotypes; when $f = 0$, noted as SVP = 2, it is referred to as environment-focused partitioning, and the biplot is most suitable for visualizing the genetic correlation among environments (Yan, 2002).

A GGE biplot is generated by plotting $\lambda_1^f \xi_{i1}$ and $\lambda_1^{1-f} \eta_{j1}$ against $\lambda_2^f \xi_{i2}$ and $\lambda_2^{1-f} \eta_{j2}$, respectively. Although this type of biplot has been used previously in MET data analysis (e.g. Cooper *et al.*, 1997), methods for the utilization of the information contained in a biplot to its fullest extent became available in the late 1990s (Yan, 1999; Yan *et al.*, 2000).

Singular value decomposition of within-environment standard deviation-scaled data

The second model that can be used to generate a GGE biplot is:

$$(Y_{ij} - \overline{Y_j})/s_j = \lambda_1 \xi_{i1} \eta_{j1} + \lambda_2 \xi_{i2} \eta_{j2} + \varepsilon_{ij} \quad \text{(Eqn 10.2)}$$

where s_j is the standard deviation for genotype means for environment j, and all other parameters are the same as in Eqn 10.1. This model removes the units of the data and assumes an equal ability of all environments to discriminate among genotypes, which may be an undesirable property for genotype–environment data analysis. It is useful for analysing genotype–trait data, however, in which different traits use different units. There are other types of GGE biplots, depending on how the data are scaled (Yan and Holland, 2010; Yan, 2014).

Biplot Analysis of Multi-environment Trial Data: An Example

This section provides an example of biplot analysis of MET data using the 1993 Ontario winter-wheat performance trial data. Efforts will be made to demonstrate how a GGE biplot can be used to address the four major utilities of MET data analysis.

The steps in biplot analysis

The sample dataset is presented in Table 10.1, which contains the mean yield of 18 winter-wheat genotypes tested in nine Ontario locations in 1993. The trials were replicated four to six times at each location, but we present only the mean data for the purpose of illustration. Generating a GGE biplot based on Eqn 10.1 involves the following steps:

1. Centring the data, i.e. subtracting the respective environmental means from each of the cells.
2. Subjecting the environment-centred data to SVD, which results in singular values – genotype and environment scores for each of the n principal components, n being the number of environments. SVD is a complex mathematical operation that decomposes a matrix into two component matrices using the least-squares method. Fortunately, it becomes a routine function in all major statistical analysis systems. The SAS package (SAS Institute, 1996) has an SVD function in the IML or MATRIX procedure, so that performing the SVD of a matrix takes no more than a single statement. The PRINCOMP procedure of SAS, which performs principal component analysis, gives outputs in which the singular values are tied with the genotype (row) eigenvectors.
3. Partitioning the singular value into genotype and environment scores for each of the principal components to form the PC1 and PC2 score for each genotype and each environment (Table 10.2). Theoretically, the singular value can be partitioned in any proportion, but two partitioning methods are most useful: $f = 1$ (noted as SVP = 1) for genotype evaluation and $f = 0$ (SVP = 2) for test environment evaluation (Yan, 2002).
4. Plotting the PC1 scores against the PC2 scores (Table 10.2) to generate a biplot. Biplots using other principal components are also possible. The plotting can be done using a spreadsheet, but the abscissa and ordinate must be drawn to scale.
5. Labelling the biplot with the genotype and environment names, which can be a very tedious job.
6. Adding supplementary lines to facilitate visualization and interpretation of the biplot.

As can be seen, although the biplot is an elegant tool for visualizing MET data, the process is tedious, if not difficult, even for well-trained biometricians. Fortunately, a Windows application, called GGEbiplot, developed by Yan (2001, 2014), is available, which has fully automated the biplot analysis process. All biplots presented on the following pages of the chapter are the direct outputs of this software. In these biplots, the genotypes are labelled with lower-case letters and the environments with upper-case letters.

Visualizing the performance of different genotypes in a given environment

This is a direct application of the biplot theory described in Fig. 10.1 and associated descriptions. To visualize the performance of different

Table 10.1. Yield data (t ha^{-1}) of 18 genotypes (Entries; Column 1) in nine environments (Columns 2 to 10) and mean across all environments (Column 11).

Entries	BH93	EA93	HW93	IN93	KE93	NN93	OA93	RN93	WP93	Mean
Ann	4.5	4.2	2.8	3.1	5.9	4.5	4.4	4.0	2.7	4.0
Ari	4.4	4.8	2.9	3.5	5.7	5.2	5.0	4.4	2.9	4.3
Aug	4.7	4.6	3.1	3.5	6.1	5.0	4.7	3.9	2.6	4.2
Cas	4.7	4.7	3.4	3.9	6.2	5.3	4.2	4.9	3.5	4.5
Del	4.4	4.6	3.5	3.8	5.8	5.4	5.1	4.1	2.8	4.4
Dia	5.2	4.5	3.0	3.8	6.6	5.0	4.0	4.3	2.8	4.3
Ena	3.4	4.2	2.7	3.2	5.3	4.3	4.2	4.1	2.0	3.7
Fun	4.9	4.7	4.4	4.0	5.5	5.8	4.2	5.1	3.6	4.7
Ham	5.0	4.7	3.5	3.4	6.0	4.9	5.0	4.5	2.9	4.4
Har	5.2	4.7	3.6	3.8	5.9	5.3	3.9	4.5	3.3	4.5
Kar	4.3	4.5	2.8	3.4	6.1	5.3	4.9	4.1	3.1	4.3
Kat	3.2	3.0	2.4	2.4	4.2	4.3	3.4	4.1	2.1	3.2
Luc	4.1	3.9	2.3	3.7	4.6	5.1	2.6	5.0	2.9	3.8
M12	3.3	3.9	2.4	2.8	4.6	5.1	3.3	3.9	2.6	3.5
Reb	4.4	4.7	3.7	3.6	6.2	5.1	3.9	4.2	2.9	4.3
Ron	4.9	4.7	3.0	3.9	6.1	5.3	4.3	4.3	3.0	4.4
Rub	3.8	5.0	3.4	3.4	4.8	5.3	4.3	4.9	3.4	4.2
Zav	4.2	4.7	3.6	3.9	6.6	4.8	5.0	4.4	3.1	4.5

Table 10.2. PC1 and PC2 scores for each genotype and each environment used in constructing the GGE biplot using Eqn 10.1b.

Genotypes	PC1	PC2	Environments	PC1	PC2
Ann	–0.13	–0.55	BH93	1.03	0.19
Ari	0.16	–0.27	EA93	0.82	0.07
Aug	0.18	–0.51	HW93	0.82	0.26
Cas	0.40	0.39	IN93	0.71	0.29
Del	0.30	–0.29	KE93	1.18	–0.50
Dia	0.30	–0.09	NN93	0.47	0.51
Ena	–0.56	–0.63	OA93	0.91	–0.78
Fun	0.48	0.98	RN93	0.20	0.58
Ham	0.37	–0.28	WP93	0.57	0.48
Har	0.35	0.48			
Kar	0.16	–0.40			
Kat	–1.26	–0.22			
Luc	–0.69	1.07			
M12	–0.88	0.13			
Reb	0.19	0.07			
Ron	0.29	0.06			
Rub	–0.10	0.53			
Zav	0.45	–0.46			

genotypes in a given environment, say, BH93, draw a line that passes through the biplot origin and the marker of BH93; this may be called the BH93 axis. The genotypes will be ranked according to their projections on to the BH93 axis (Fig. 10.2). Thus, the yield order of the genotypes in BH93 is: Kat < M12 < Ena < Luc < Ann < ... < Har ≈ Cas < Fun. The line passing through the biplot origin and perpendicular to the BH93 axis separates genotypes that yielded below the mean (Kat, M12, Ena, Luc and Ann) from genotypes that yielded above the mean (all other genotypes) in BH93.

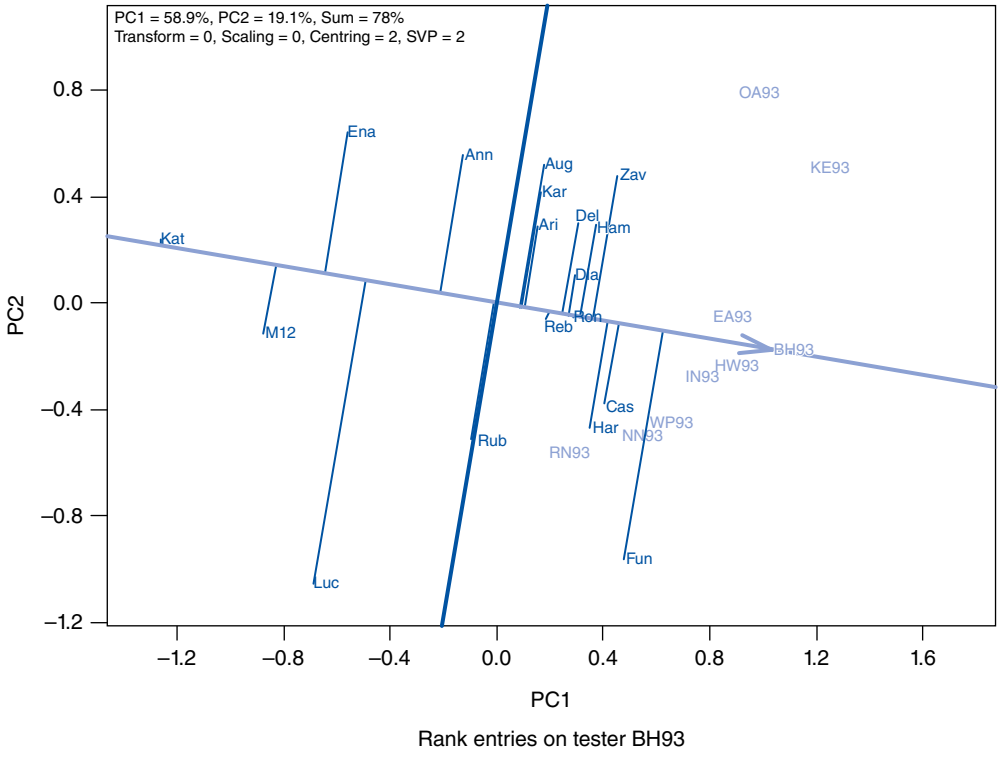


Fig. 10.2. Ranking of the genotypes based on their performance in environment BH93.

Visualizing the relative adaptation of a given genotype in different environments

Analogous to the above, to visualize the relative performance of a given genotype, say, Rub, in different environments, draw a line that passes through the biplot origin and the marker of Rub, which may be called the Rub axis. The environments are ranked along the Rub axis in the direction towards the marker of Rub (Fig. 10.3). Thus, the order of performance of Rub in different environments is: RN93 > NN93 > WP93 > IN93 > BH93 > EA93 > KE93 > OA93. The line passing through the biplot origin and perpendicular to the Rub axis separates environments in which Rub yielded below the mean (OA93, KE93 and EA93) from environments in which Rub yielded above the mean (all other environments, except BH93). Environment BH93 was right on the perpendicular line, implying that Rub yielded near the overall mean in BH93.

Visual comparison of two genotypes in different environments

Biplot comparison of two genotypes is an extension of the basic biplot principle. To compare two genotypes, connect the two genotypes to be compared, say, Aug and Rub, with a straight line (called a connector line) and draw a line that is perpendicular to the connector line and to pass through the biplot origin (Fig. 10.4). This perpendicular line separates environments where Aug yielded better than Rub from environments where Rub yielded better than Aug. Thus, Fig. 10.4 reveals that Aug yielded higher than Rub in OA93, KE93, EA93 and BH93, and Rub yielded higher than Aug in the other five environments. Based on the basic principle of biplot geometry described earlier, the two genotypes would yield exactly the same in environments whose markers fall on the perpendicular line. If all environments fall on the same side of the perpendicular line, the genotype with the environments on its

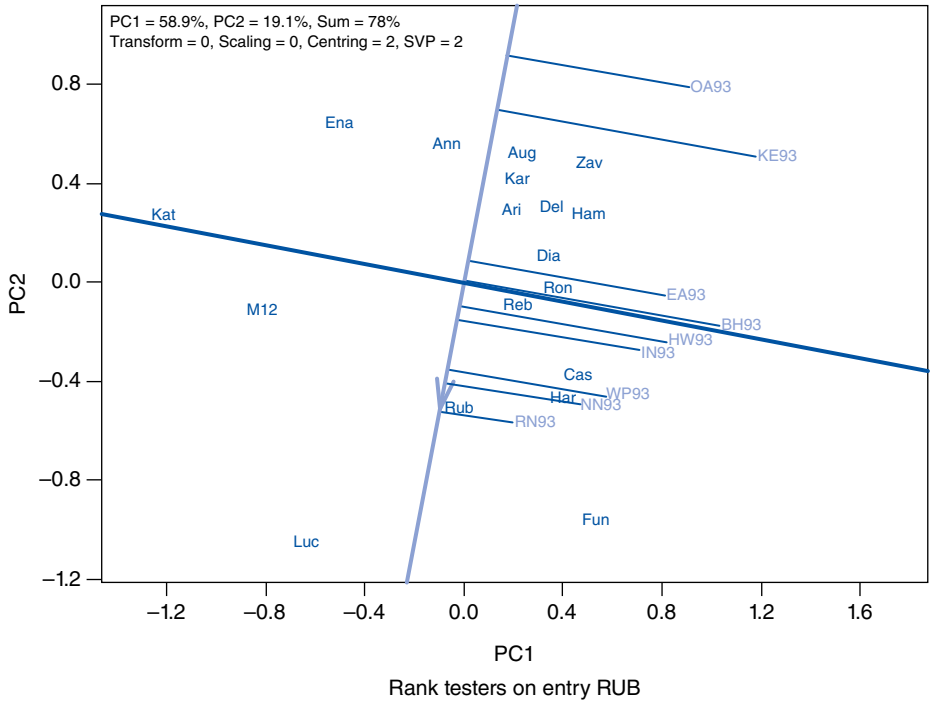


Fig. 10.3. Ranking of the environments based on the relative performance of genotype Rub.

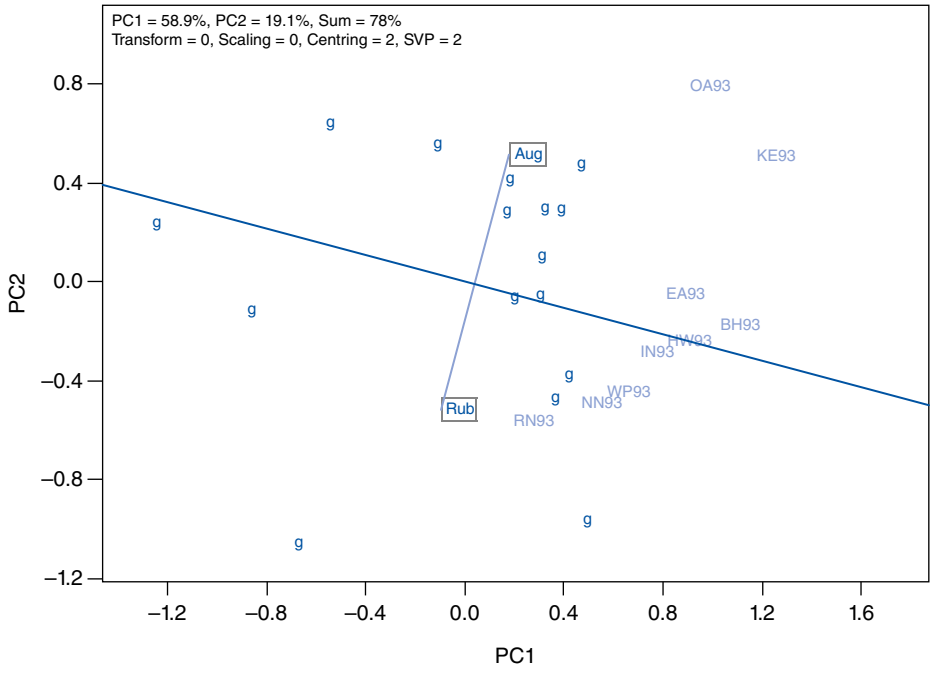


Fig. 10.4. Comparison of the two genotypes Aug and Rub in different environments.

side would yield better than the other genotype in all environments. If the two genotypes are spatially close, they are likely to have yielded similarly in all or most of the environments.

Visual identification of the best genotype(s) for each environment

A further extended application of the biplot geometry is to visually identify the highest-yielding genotypes for each of the environments in a single step. For this purpose, the genotypes that are located far away from the biplot origin are connected with straight lines so that a polygon or vertex hull is formed with all other genotypes contained within the vertex hull (Fig. 10.5). The vertex genotypes in our example are Fun, Zav, Ena, Kat and Luc. These genotypes are the most responsive genotypes; they are either the best or the poorest genotypes in some or all of the environments. Perpendicular lines to the sides of the vertex hull are drawn, starting from the biplot origin, to divide the biplot into five sectors or quadrants, each having a vertex genotype. The beauty of Fig. 10.5 is that the vertex genotype for each quadrant is, according to the biplot approximation, the one that gave the highest yield for the environments that fall within that quadrant. Thus, genotype Fun gave the highest yield in environments RN93, NN93, WP93, IN93, HW93, BH93 and EA93, and genotype Zav gave the highest yield in environments OA93 and KE93. The other vertex genotypes, i.e. Ena, Kat and Luc, did not give the highest yield in any of the environments. Actually, they were the poorest genotypes in some or all of the environments.

Now we explain why the above statements are valid. According to the section 'Visual comparison of two genotypes in different environments', the line perpendicular to the polygon side that connects genotypes Luc and Fun facilitates the comparison between Luc and Fun; Fun yielded higher than Luc in all environments because all environments are on the side of Fun. Likewise, the line perpendicular to the

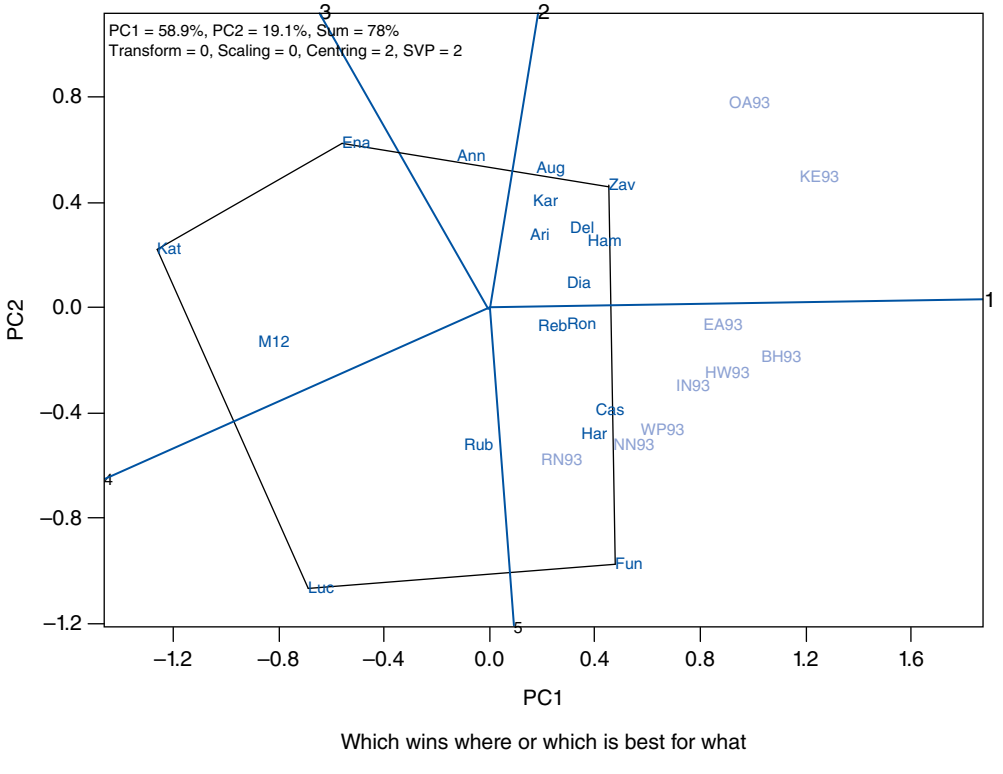


Fig. 10.5. The polygon view of the GGE biplot indicating the best genotype(s) in each environment and groups of environments.

polygon side that connects genotypes Zav and Fun facilitates the comparison between Zav and Fun; Fun yielded higher than Zav in seven environments that fall into the Fun sector because they are on the side of Fun. Within the Fun sector, Fun has the longest vector (distance from biplot origin to the marker of a genotype); it therefore gave higher yields than other genotypes in these seven environments, for reasons discussed in the section 'Visualizing the performance of different genotypes in a given environment'. Collectively, Fun gave the highest yield in environments that fell in its sector. Using the same reasoning, Zav was the best genotype in environments KE93 and OA93.

Visualizing groups of environments

Another utility of Fig. 10.5 is that the environments are grouped based on the best genotypes. We have two groups of environments: KE93 and OA93 as one group, with Zav as the highest-yielding genotype, and the other seven environments as another group, with Fun as the highest-yielding genotype.

The environment groups suggest different mega-environments. In our example, KE93 and OA93 represent eastern Ontario and the other environments represent western and southern Ontario. The hypothesis that eastern Ontario is a different mega-environment from the rest of Ontario for winter-wheat production was tested and confirmed using 1989–2000 Ontario winter-wheat performance trial data (Yan, 1999). The genotype by environmental group interaction explained 80% of the total GEI (Yan, 1999).

Visualizing the mean performance and stability of genotypes

Once mega-environments are defined, cultivar selection should be specific to individual mega-environments. For a given mega-environment, genotypes are evaluated based on mean performance (such as mean yield) and stability across environments. Assuming that the nine environments in our example belong to a single mega-environment, a 'mean' environment can be defined in the biplot, using the mean-environment PC1 and PC2 scores of all environments. The mean yield of the genotypes can then be approximated by the nominal yields of the genotypes in that mean environment.

In Fig. 10.6, a line is drawn that passes through the biplot origin and the mean environment, which is marked by a small circle. This line will be called the mean-environment axis. Another line is drawn that passes through the biplot origin and is perpendicular to the mean-environment axis. These two lines constitute 'the mean-environment coordination'.

The projections of the genotypes on to the mean-environment axis approximate the mean yield of the genotypes. Thus, the mean yield of the genotypes is in the following order: Fun > Cas ≈ Har > ... > Rub > Ann > Luc > Ena > M12 > Kat. This order is highly consistent with the actual mean yield of the genotypes (Table 10.1). The parallel lines in Fig. 10.6 facilitate ranking of the genotypes based on their predicated mean yield. Since the biplot contains both G and GEI and since the two axes of the mean-environment coordination are orthogonal, if projections of the genotypes on to the mean-environment axis approximate the mean yield of the genotypes, projections of the genotypes on to the perpendicular axis must approximate the GEI associated with the genotypes. The longer the projection of a genotype, regardless of direction, the greater the GEI associated with the genotype, which is a measure of variability or instability of the genotype across environments. Thus, the performances of genotypes Luc and Fun are highly variable (less stable), whereas genotypes Ron and Reb are highly stable.

It should be pointed out that high stability is not necessarily a positive thing per se. High stability is desirable only when associated with a high mean yield. A genotype with high stability is highly undesirable if it is associated with a low mean yield; it is simply a genotype that is consistently poor. It is even less desirable than genotypes with poor stability.

An ideal genotype is one that has both high mean yield and high stability. The centre of the concentric circles in Fig. 10.7a represents the position of the 'ideal' genotype, which is defined by a projection on to the mean-environment axis that equals the longest vector of the genotypes that had above-average mean yield and by a zero

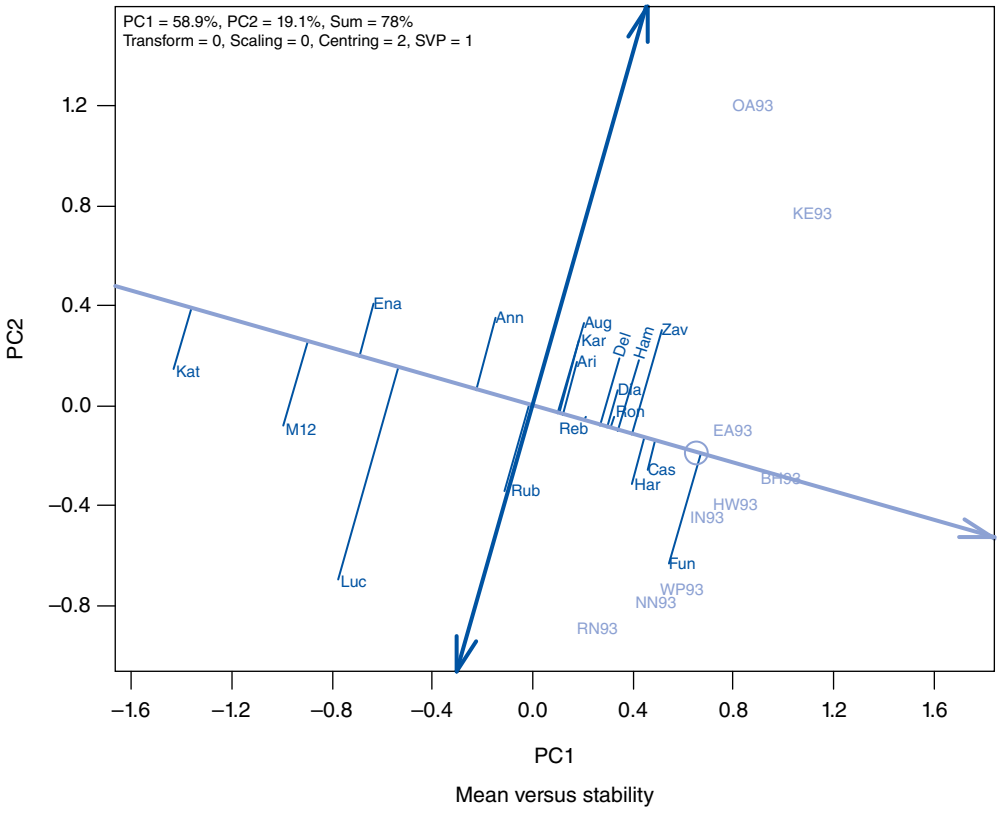


Fig. 10.6. The mean-environment coordination showing the mean yield and stability of each of the genotypes.

projection on to the perpendicular line (zero variability across environments). A genotype is more desirable if it is closer to the 'ideal' genotype. Thus, genotypes Cas and Har are equally desirable as genotype Fun, even though the latter had the highest mean yield. The low-yielding genotypes Kat, M12, Luc, Ena and Ann, are, of course, undesirable because they are far away from the 'ideal' genotype.

Visualizing the discriminating ability and representativeness of environments

Although METs are conducted primarily for genotype evaluation, they can also be used in evaluating environments. An ideal environment should be highly differentiating of the genotypes and at the same time representative of the target environment. Assuming that the test environments used in the MET are representative samples of the target environment, the ideal environment should be located on the mean-environment axis. The centre of the concentric circles represents the ideal environment, which has the longest vector of the test environments that had positive projections on to the mean environment axis (Fig. 10.7b). An environment is more desirable if it is closer to the 'ideal' environment. Therefore, BH93, EA93, HW93 and IN93 were relatively desirable test environments, whereas OA93 and RN93 were relatively undesirable test environments.

Strength of the GGE-biplot approach

The GGE-biplot approach graphically displays genotype main effect and GEI of an MET, which are the two parts of yield variation that are pertinent to genotype evaluation and mega-environment identification. Assuming that the GGE of an MET is sufficiently

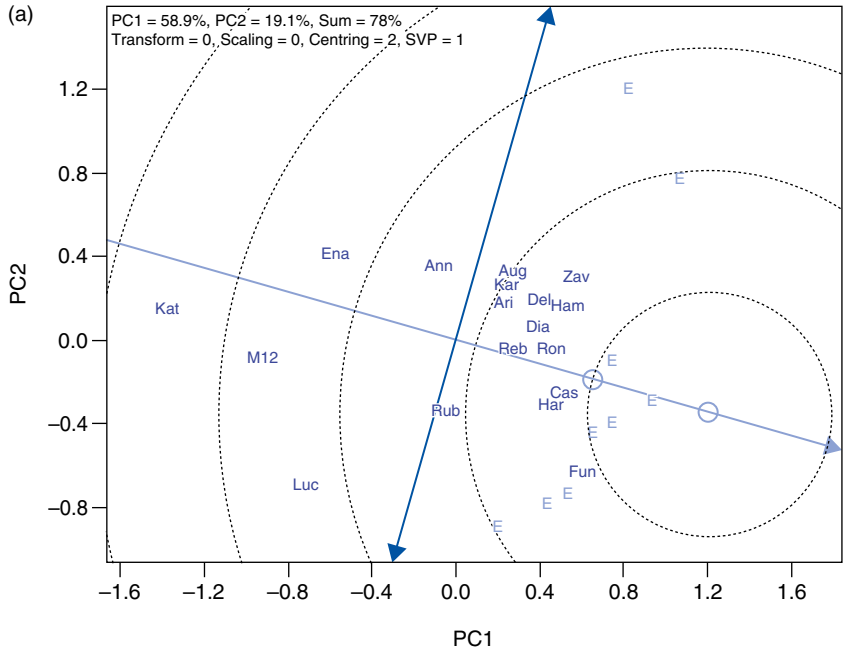


Fig. 10.7. Comprehensive evaluation of genotypes and environments. (a) Comparison of genotypes with the 'ideal' genotype for both mean yield and stability. (b) Comparison of environments with the 'ideal' environment based on both discriminating ability and representativeness of the target environment.

approximated by the first two principal components, all individual genotype–environment relationships in the MET should be displayed by the GGE biplot. Such a biplot graphically addresses three of the four utilities of MET data analysis listed in the introduction of this chapter, namely: (i) investigating possible mega-environment differentiation in the target environment; (ii) selecting superior genotypes for individual mega-environments; and (iii) selecting better test environments. In addition, the GGE biplot also facilitates pairwise genotype comparisons. The GGE biplot does not directly address the fourth utility of the MET data analysis, i.e. understanding the causes of GEI. To fulfil this task, information other than yield per se is necessary. Once such information is available, the genotype and environment scores can be related to genotypic and environmental factors, so that the observed GEIs can be explained as interactions between genotypic factors and environmental factors (Yan and Hunt, 2001). Therefore, the GGE biplot is an effective and versatile approach for MET data analysis.

Constraints of the GGE-biplot approach

All methods have their limitations. The limitations of the GGE biplot lie in four aspects. First, it requires balanced data; second, it may explain only a small portion of the total GGE; third, it lacks a measure of uncertainty; and fourth, although elegant, GGE-biplot analysis is tedious to perform using conventional tools. Now that the GGEbiplot software and other software packages are available, the fourth constraint is no longer an issue. Once the data are properly arranged, all functions are just a 'mouse-click' away. All the figures presented in this chapter, along with many other options, are the direct outputs of GGEbiplot.

Although quite common, unbalanced MET data are really not a problem related to the GGE-biplot approach; they are a problem related to the experimental design and execution, which creates problems for all kinds of analyses. The milling value imputing method reported in Yan (2013) can solve this problem.

The GGE biplot may explain only a small proportion of the GGE when the genotype main effect is considerably smaller than the GEI and when the GEI pattern is complex. In such cases, the GGE biplot, consisting of PC1 and PC2, may not be sufficient to explain the GGE, even though the most important pattern of the MET is already displayed. To remedy this problem, the GGEbiplot software offers options for viewing biplots of PC3 versus PC4, PC5 versus PC6, etc.

Unlike conventional approaches, which allow calculation of probability for a particular hypothesis, the GGE-biplot approach does not have a measure of uncertainty. Therefore, the GGE biplot is better used as a hypothesis generator rather than as a decision maker (Yan *et al.*, 2001), and hypotheses based on biplots should be tested using conventional statistical methods. For example, biplots based on individual years of Ontario winter-wheat performance trials suggested that eastern Ontario sites and other sites of Ontario belong to different mega-environments, and this hypothesis was tested and confirmed by variance component analysis (Yan, 1999). Sometimes, the biplot distance of two genotypes, relative to the biplot size, may be sufficiently informative about the significance of the difference between two genotypes or two environments.

Other applications of the GGE-biplot approach

The GGE-biplot methodology was developed for MET data analysis. It is a generic method, however. It has been successfully used in analysing genotype–trait data (Yan and Rajcan, 2002), diallel-cross data (Yan and Hunt, 2002) and host genotype–pathogen race data (Yan and Falk, 2002). The GGE-biplot methodology and the GGE biplot software described in this chapter should thus be useful for the graphical presentation of all types of two-way data that conform to an entry–tester data structure.

New Developments Regarding GGE-biplot Analysis Since 2002

GGE biplot analysis has been widely adopted by the research community since the first edition of this book and the publication of a textbook by Yan and Kang (2003), which described in detail

the various aspects of GGEbiplot software. Subsequently, Yan (2014) has expounded on the use of GGE-biplot analyses for MET data management and analysis in another book. Currently, more than 81,000 webpages and 4500 scholarly papers contain the word 'GGE biplot'. This section briefly summarizes the progress that has been made since then.

Heritability-enriched GGE biplot for test environment evaluation

A GGE biplot can be used to assess the representativeness and discriminating ability of the test environment (Fig. 10.7a). According to Allen *et al.* (1978), the proper measure of the value of a test environment is $r\sqrt{H}$, where r is the correlation between genotypic performance in a test environment and that in the target environment (the representativeness), and H is the heritability in the test environment (the discriminating ability). For a GGE biplot to display these two aspects, a heritability-adjusted (HA) GGE biplot was developed (Yan and Holland, 2010). This GGE biplot is based on the model in Eqn 10.2, except that the standardized genotype-by-environment two-way table is multiplied by $\sqrt{H_j}$ for each environment before submitting it to singular value decomposition. The so-called 'HA-GGE biplot' has the interpretation that the cosine of the angle between a test environment and the average environment indicates its representativeness and the vector length of an environment indicates its discriminating ability. More systematic descriptions of model selection in biplot analysis of a two-way table can be found in Yan (2014).

GGE-GGL biplot to identify repeatable GEI

The 'which-won-where' pattern in a GGE biplot (Fig. 10.5) was suggested to indicate mega-environments. However, meaningful mega-environment delineation must be based on repeatable 'which-won-where' patterns. This means that data from multi-year, multi-location trials are needed for mega-environment delineation. Often different sets of genotypes are tested in different years; thus, most studies have adopted the 'analyse yearly, summarize across years' strategy (DeLacy *et al.*, 1996). However, summarizing across years is very difficult (Yan *et al.*, 2000; Casanoves *et al.*, 2005; Mohammadi *et al.*, 2009; Munaro *et al.*, 2014). The missing value estimation method of Yan (2013) allowed unbalanced multi-year data to be presented in a single GGE biplot, treating each location–year combination as an environment. On such a GGE biplot, the locations are also labelled, and the placement of a location is determined by the mean coordinates of all trials at the location. This biplot is, therefore, referred to as a 'GGE-GGL biplot.' Repeatable GEI patterns and, therefore, the presence of different mega-environments, are indicated when the locations fall into clear groups (Yan, 2014, 2015, 2016); otherwise it indicates that the GEI is dominated by unrepeatable GE and the locations cannot be divided into meaningful mega-environments.

LG biplot for mega-environments analysis

The LG biplot (Yan, 2019) is a new development after the GGE-GGL biplot. While the GGE-GGL biplot method requires some common genotypes to be tested across years to impute missing values, the LG biplot does not require this. It is, therefore, more flexible. Using this method, all crop variety trial data can be utilized in mega-environment analysis. The LG biplot is a biplot that displays the correlations among test locations in each of the years in a multi-location, multi-year dataset. It assumes that the genotypes tested each year are random samples of a genotype population. Like the GGE-GGL biplot, the placement of a location in the LG biplot is determined by the mean coordinates of all trials at the location. Repeatable GEI patterns and, therefore, the presence of different mega-environments, are indicated when the locations fall into clear groups.

GYT biplot for genotype evaluation based on multiple traits

Final genotype evaluation and decision making in plant breeding have to consider multiple traits.

A genotype-by-trait (GT) biplot can display a genotype by trait two-way table and allows visualizing the strengths and weaknesses of the genotypes (Yan and Rajcan, 2002). However, it cannot be used in making selection decisions due to strong negative correlations among traits. Some traits are so measured that a smaller value means more desirable (e.g. lodging or disease score). The newly developed GYT (genotype by yield*trait) biplot has solved this problem. It graphically ranks the genotypes for their overall superiority as well as their trait profiles (Yan and Frégeau-Reid, 2018). The concept underlying GYT biplot analysis is that yield has an unparalleled importance in genotype evaluation and that the economic value of the level of a trait (other than yield) is dependent on the level of yield with which it is associated. For example, superior lodging resistance has little economic value when associated with low yield but is increasingly more valuable when associated with higher yield.

The GYT biplot analysis includes the following steps. First, transform the genotype by trait two-way table into a genotype by yield*trait (GYT) two-way table, in which a higher value is always more desirable. Second, standardize this table by each yield–trait combination. Third, calculate the mean across yield–trait combinations for each genotype, which can be called the GYT index, which is a measure of the overall superiority of the genotypes in combining yield with other traits. Finally, display the standardized GYT table in a GYT biplot so that the overall superiority, and the strengths and weaknesses of the genotypes can be visualized. A most recent development of this methodology (Yan *et al.*, 2019) is to allow the yield–trait combinations to have different weights so that the GYT index can be used as the sole criterion in making selection decisions for a given target environment and end-use. Weighting is not necessary if the decision is made on the GYT biplot, i.e. on both overall superiority and trait profiles. The new developments in the use of biplot methodology have expanded the toolkit of plant breeders and agronomists in analysing MET data.

Acknowledgements

We acknowledge Dr Rich Zobel and Mr Hugh Gauch for their stimulating suggestions and critiques during the development of the GGE-biplot methodology. Drs Paul Cornelius and Jose Crossa are also acknowledged for their encouragement and editing of our first paper on the GGE-biplot methodology.

References

Allen, F.L., Comstock, R.E. and Rasmusson, D.C. (1978) Optimal environments for yield testing. *Crop Science* 18, 747–751.

Casanoves, F., Baldessari, J. and Balzarini, M. (2005) Evaluation of multienvironment trials of peanut cultivars. *Crop Science* 45(1), 18–26.

Cooper, M., Stucker, R.E., DeLacy, I.H. and Harch, B.D. (1997) Wheat breeding nurseries, target environments, and indirect selection for grain yield. *Crop Science* 37, 1168–1176.

DeLacy, I.H., Basford, K.E., Cooper, M., Bull, J.K., and McLaren, C.G. (1996) Analysis of multi-environment trials–an historical perspective. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK, pp. 39–124.

Finlay, K.W. and Wilkinson, G.N. (1963) The analysis of adaptation in a plant breeding program. *Australian Journal of Agricultural Research* 14, 742–754.

Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.

Gauch, H.G. and Zobel, R.W. (1996) AMMI analysis of yield trials. In: Kang, M.S. and Gauch, H.G. (eds) *Genotype-by-Environment Interaction*. CRC Press, Boca Raton, Florida, pp. 1–40.

Gauch, H.G. and Zobel, R.W. (1997) Identifying mega-environments and targeting genotypes. *Crop Science* 37, 311–326.

Kang, M.S. (1993) Simultaneous selection for yield and stability in crop performance trials: Consequences for growers. *Agronomy Journal* 85(3), 754–757. DOI: 10.2134/agronj1993.00021962008500030042x.

Kang, M.S. (1998) Using genotype by environment interaction for crop cultivar development. *Advances in Agronomy* 62, 199–252.
Lin, C.S. and Binns, M.R. (1994) Concepts and methods for analysis regional trial data for cultivar and location selection. *Plant Breeding Reviews* 11, 271–297.
Lin, C.S., Binns, M.R. and Lefkovitch, L.P. (1986) Stability analysis: Where do we stand? *Crop Science* 26, 894–900.
Mohammadi, R., Amri, A. and Ansari, Y. (2009) Biplot analysis of rainfed barley multienvironment trials in Iran. *Agronomy Journal* 101, 789–796.
Munaro, L.B., Benin, G., Marchioro, V.S., de Assis Franco, F., Silva, R.R., *et al*. (2014). Brazilian spring wheat homogeneous adaptation regions can be dissected in major megaenvironments. *Crop Science* 54(4), 1374–1383.
SAS Institute Inc. (1996) *Version 6, SAS/STAT User's Guide*. SAS Institute, Cary, North Carolina.
Yan, W. (1999) Methodology of cultivar evaluation based on yield trial data – with special reference to winter wheat in Ontario. PhD dissertation, University of Guelph, Guelph, Ontario, Canada.
Yan, W. (2001) GGEbiplot – a Windows application for graphical analysis of multi-environment trial data and other types of two-way data. *Agronomy Journal* 93(5), 1111.
Yan, W. (2002) Singular-value partitioning in biplot analysis of multienvironment trial data. *Agronomy Journal* 94(5), 990–996.
Yan, W. (2013) Biplot analysis of incomplete two-way tables. *Crop Science* 53, 48–57.
Yan, W. (2014) *Crop Variety Trials: Data Management and Analysis*. Wiley, New York.
Yan, W. (2015) Mega-environment analysis and test location evaluation based on unbalanced multiyear data. *Crop Science* 55(1), 113–122.
Yan, W. (2016) Analysis and handling of G × E in a practical breeding program. *Crop Science* 56, 2106–2118. DOI: 10.2135/cropsci2015.06.0336.
Yan, W. (2019) LG biplot: A graphical method for mega-environment investigation using existing crop variety trial data. *Scientific Reports* 9, 7130. DOI: 10.1038/s41598-019-43683-9.
Yan, W. and Falk, D.E. (2002). Biplot analysis of host-by-pathogen data. *Plant Disease* 86(12), 1396–1401.
Yan, W. and Frégeau-Reid, J. (2018) Genotype by yield*trait (GYT) biplot: A novel approach for genotype selection based on multiple traits. *Scientific reports* 8, 8242. DOI: 10.1038/s41598-018-26688-8.
Yan, W. and Holland, J.B. (2010) A heritability-adjusted GGE biplot for test environment evaluation. *Euphytica* 171(3), 355–369. DOI: 10.1007/s10681-009-0030-5.
Yan, W. and Hunt, L.A. (2001) Genetic and environmental causes of genotype by environment interaction for winter wheat yield in Ontario. *Crop Science* 41(1), 19–25.
Yan, W. and Hunt, L.A. (2002) Biplot analysis of diallel data. *Crop Science* 41(1), 21–30.
Yan, W. and Kang, M.S. (2003) *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC Press, Boca Raton, Florida. 271 pp.
Yan, W. and Rajcan, I. (2002) Biplot evaluation of test sites and trait relations of soybean in Ontario. *Crop Science* 41(1), 11–20.
Yan, W., Hunt, L.A., Sheng, Q. and Szlavnics, Z. (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science* 40(3), 597–605.
Yan, W., Cornelius, P.L., Crossa, J. and Hunt, L.A. (2001) Two types of GGE biplots for analyzing multi-environment trial data. *Crop Science* 41(3), 656–663.
Yan, W., Frégeau-Reid, J., Mountain, N. and Kobler, J. (2019) Genotype and management evaluation based on Genotype by yield*trait (GYT) analysis. *Crop Breeding, Genetics and Genomics* 1, e190002. DOI: 10.20900/cbgg20190002
Zobel, R.W., Wright, M.J. and Gauch, H.G. (1988) Statistical analysis of a yield trial. *Agronomy Journal* 80, 388–393.

11 Design and Analysis of Multi-year Field Trials for Annual Crops

Vivi N. Arief, Ian H. DeLacy and Kaye E. Basford*
The University of Queensland, Brisbane, Australia

Introduction

Elite genotypes are one of the primary goals of any plant breeding programme. These elite genotypes are identified by selecting the best performing genotypes from each breeding cycle. Selection is commonly made in stages, referred to here as selection phases, where the number of genotypes at the beginning of a breeding cycle is progressively reduced until the elite genotypes are identified at the end. At each selection phase, genotypes are grown in multi-location field trials to evaluate their relative performance for a target population of environments (TPE) (Basford and Cooper, 1998).

This multi-phase selection scheme is adopted by plant breeding programmes as a response to the presence of genotype-by-year interaction (GYI). Genotype-by-year (GY) and genotype-by-year-by-location (GYL) interactions are usually bigger than genotype-by-location interaction (GLI) (e.g. Patterson *et al.*, 1977; Cullis *et al.*, 1996a; DeLacy *et al.*, 1996; Frensham *et al.*, 1999). These interactions contribute to genotype-by-environment interaction (GEI), which reduces the effectiveness of selection in any one environment (Basford and Cooper, 1998), as genotypes are likely to have different relative performances across environmental conditions. The variance attributable to GEI is also generally higher than the genotypic variance (Bull *et al.*, 1992), making it difficult to select the best and most stable genotypes. Typically, locations are repeatable. Hence GLI can be managed though location grouping, such as mega-environments. However, years are less repeatable, so GYI cannot be managed through year grouping.

For annual field crops, each selection phase usually corresponds to a year. Therefore, selection in each phase is usually based on the genotype performance estimated from single-year multi-location field trials. Analysis of these trials has a disadvantage in that it cannot provide an estimate of GYI and genotype-by-year-by-location interaction (GYLI) and hence could lead to bias in the estimation of genotype performance. In any single-year analysis, genotypic variance (VG) and genotype-by-location variance (VGL) are likely to be overestimated. The estimate of VG is confounded with the estimate of genotype-by-year variance (VGY) and the estimate of VGL is confounded with the estimate of genotype-by-year-by-location variance (VGYL) (Nyquist and Baker, 1991; Holland and Nyquist, 2010; Arief *et al.*, 2015, 2019).

Analysis of multi-year data can eliminate these confounding effects (Arief *et al.*, 2019). Moreover, the change from a single-year to a

* Email: k.e.basford@uq.edu.au

© CAB International, 2020.

multi-year analysis requires a minimal change in resources because multi-year data are readily available in most plant breeding programmes. In addition to a better estimate of genotype performance (Arief *et al.*, 2019), analysis of multi-year data provides a reasonable estimate of VG, VGL, VGY, VGYL and the residual variance (VR) (e.g. DeLacy *et al.*, 2010; Barrero Farfan *et al.*, 2013; Arief *et al.*, 2015). These estimated variance components are useful for optimizing and redesigning the field trials in a plant breeding programme (Arief *et al.*, 2015).

In field trials, a field is a single experiment. In each field, genotypes targeted for the same TPE are usually grown together regardless of their breeding cycle. Therefore, a field trial in any single year often contains a mixture of genotypes across breeding cycles and selection phases. Common checks are usually included to provide connection among breeding cycles or selection phases (Piepho *et al.*, 2006). As these genotypes are likely to be grown under the same management conditions (as they are within the same field), it is recommended that an experimental design is applied to the whole field to minimize the estimation of residual error within the field (Basford *et al.*, 1996; Williams and John, 1996; Federer, 2005). In multi-environment (i.e. multi-location and multi-year) trials, it is also recommended that the design is across locations and across years so as to optimize the overall efficiency of the multi-environment trials (METs) (e.g. Sprague and Federer, 1951; Arief *et al.*, 2015).

This chapter provides a brief discussion on the design and analysis (prediction and interpretation) of field trials, focusing on multi-environment yield trials for annual crops.

Design of Field Trials

In a single year, genotypes from a plant breeding programme are usually tested in many trials across many fields in many locations, i.e. in multi-location trials. These multi-location trials are conducted to obtain an estimate of genotype performance across a range of locations representing the TPE. However, as each field is likely to be heterogeneous, an experimental design should be used to account for this heterogeneity. An experimental design should also be applied to account for heterogeneity across fields, both within and across years. Therefore, there are two levels of designs for field trials: an experimental design applied to each field (within-field design) and an experimental design applied across fields (across-fields design).

Within-field design

There are three principles of experimental designs: randomization, replication and blocking (Cochran and Cox, 1957). Randomization reduces the potential bias attributable to systematic placing of the genotypes. Thus, a separate randomization should be applied in each field. Replication is used to estimate the random variance that cannot be controlled by the experimental design. Blocking is used to control for some confounding factors that could affect the response variable and could not be separated from the explanatory variable in any other way.

An agricultural experiment is traditionally designed using a balanced complete block design, such as randomized complete block or Latin square designs. These designs are 'complete', as each block contains all treatments, and 'balanced' as treatments occur in equal frequency. As it is assumed that within-block variation is smaller than among-blocks variation (Cochran and Cox, 1957), a complete block design is generally not suitable for a plant breeding field trial where a large number of treatments (genotypes) are being tested.

Consequently, an incomplete block design is more appropriate (Cochran and Cox, 1957). One relevant class of this type of design is called a resolvable incomplete block design, i.e. the blocks must be capable of arrangement in complete replications (e.g. Yates, 1937; Patterson and Williams, 1976; Patterson and Robinson, 1989; Williams and John, 1996; Williams *et al.*, 2006).

Lattices and alpha-designs

The lattice square designs, introduced by Yates (1937), are balanced, efficient and easy to analyse (Patterson and Robinson, 1989) but are only available for a limited number of treatments and blocks. Patterson and Williams (1976) introduced alpha-designs, which have no limitation on block size, other than the unavoidable

constraint that the block size k must be a factor of the number of entries v. These designs are used extensively in plant breeding because of their flexibility regarding the number of entries (genotypes), the size of the incomplete block and their ability to provide good error control (Yau, 1997).

Row–column designs

In a plant breeding field trial, plots are usually laid out in rows and columns to give a compact block (Patterson and Robinson, 1989). This two-dimensional arrangement of the plots enables the use of a resolvable row-column design (Williams *et al.*, 2006). These designs place fewer restrictions on the number of genotypes and replications than lattices and alpha-designs (Patterson and Robinson, 1989). Kempton *et al.* (1994) demonstrated greater efficiency from the use of two-dimensional blocking structures in the analysis of 244 UK cereal trials. Williams and John (1996) strongly recommended the use of row–column designs, wherever the plots are laid out in a rectangular grid. The CycDesign software (VSN International; http://www.vsni.co.uk), the R package DiGGer (Coombes, 2009) and a web-based application DeltaGen (Jahufer and Luo, 2018) provide facilities for constructing efficient resolvable row–column designs.

Latinized rows and columns

The row–column design minimizes the occurrences of treatments, here genotypes, in the same row or column within a replication and also minimizes the chance of the same pair of treatments occurring next to each other in different replications (John and Williams, 1995). In a field trial, replications are usually laid out next to each other along the columns (i.e. long columns, Fig. 11.1a) or the rows (i.e. long rows, Fig. 11.1b).

A further restriction can be applied in a row–column design to ensure that a treatment does not appear more than once in a long column or a long row. This restriction is known as a latinized row–column design (John and Williams, 1995).

It is common in a plant breeding field trial to test genotypes from multiple sets in the same field. These sets can be based on breeding cycles, selection phases, breeders' groups or breeding objectives. Because these genotypes are grown in the same field, it is reasonable to assume that they are grown under the same management conditions. Ideally, an experimental design should be applied to the whole field using long rows and long columns to minimize the error within the field (Basford *et al.*, 1996; Williams and John, 1996; Federer, 2005). However, in practice, an experimental design is applied to each set (Arief *et al.*, 2019). As each set is likely to consist of different genotypes, except for few overlapping genotypes, latinization across long rows or long columns within a field might not be required.

Repeated checks

A balanced design, where all genotypes have the same number of replications, often cannot be accommodated in plant breeding field trials when supplies of seed and other resources are limited. However, common checks are often included to provide a connection across sets (Piepho *et al.*, 2006; Arief *et al.*, 2019). These common checks are usually replicated and can be used to provide an estimate of field residual variance (Arief *et al.*, 2019).

Across-fields design

Multi-environment trials are conducted to obtain an estimate of genotype performance across

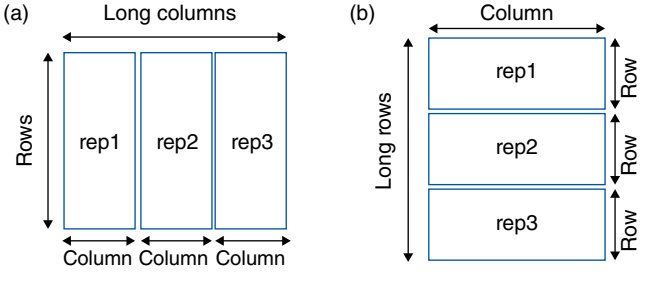


Fig. 11.1. Field layout with rows and columns. (a) Replications are laid out along the long columns. (b) Replications are laid out along the long rows.

locations and years. There are three criteria commonly used to measure the trialling efficiency of field trials in a plant breeding programme: genetic repeatability (also known as broad-sense line-mean heritability), acceptance probability and potential gain. Each of these derived statistics highlights a different feature of trialling efficiency. Genetic repeatability is a measure of the accuracy of the field testing in estimating the genotype performance (Fehr, 1987); the acceptance probability is a measure of risk of rejecting good genotypes (Patterson *et al.*, 1977); and the potential gain is a measure of the potential improvement in the mean of selected genotypes (Talbot, 1997). Each of these can be used to optimize resource allocation in terms of the number of genotypes, replications, locations and years (e.g. Sprague and Federer, 1951; Patterson *et al.*, 1977; Talbot, 1997) and can be combined to redesign a field trial (Arief *et al.*, 2015).

Within years

Plant breeding field trials are often conducted as unreplicated trials. These unreplicated trials are necessary, especially in early generation trials because of the limitation on seed availability (Kempton, 1984). For a fixed total number of plots, unreplicated trials also enable more genotypes to be tested (Kempton, 1984). Some plots are assigned for replicated checks (i.e. check plots) to obtain an estimate of the field's residual variance. This approach can lead to two potential problems: (i) the residual variance estimated from replicated checks might not be relevant to the test genotypes, and (ii) the performance of the repeated checks is estimated with higher precision than that of the test genotypes.

Cullis *et al.* (2006) proposed a new class of design referred to as '*p*-rep' designs, where *p* is the ratio of check plots to test plots. In these designs, some or all of the check plots are replaced with some of the test genotypes. These test genotypes could be randomly selected, could be determined by the seed availability or decided by the breeder (Cullis *et al.*, 2006). Increasing the number of fields has been shown to produce greater improvement in trialling efficiency than increasing the number of replications within a location (Fig. 11.2) (Sprague and Federer, 1951; Patterson *et al.*, 1977; Talbot, 1997). Therefore, limited replication should be more effective by being carried out across fields, rather than within a single field (Kempton, 1984). In a multi-location trial, the *p*-rep designs potentially balance the replication of the test genotypes across fields (Cullis *et al.*, 2006).

Comparison bias could potentially occur when the same pair of genotypes were tested more often across fields, especially when there is a competition effect. Latinization across fields would minimize the chance of the same pair of genotypes occurring more often across fields, and hence reduce the likelihood of comparison bias. When the genotypes are tested within sets, latinization across fields could be applied for each set.

Across years

In multi-year trials, common checks also provide a connection across years for estimating GYI and GYLI Arief *et al.*, (2015). In a multi-phase selection scheme, an additional connection across years is provided by the genotypes from each phase of selection in a breeding cycle. Therefore, in a multi-phase selection scheme, the number of common genotypes across years is usually sufficiently large to provide a reasonable estimate of GYI and GYLI (Arief *et al.*, 2019).

In contrast, the number of common genotypes across years in field trials designed for dissemination of germplasm, such as in CIMMYT's international nurseries, is somewhat limited and heavily reliant on the common checks (Arief *et al.*, 2015). In such trials, because of fixed resources, increasing the number of common checks across years will have an undesirable outcome of reducing the number of new genotypes to be tested. In this case, the use of *p*-rep designs (Cullis *et al.*, 2006) across years is potentially helpful by providing a better connection across years while maintaining the same number of new genotypes and the same amount of resources (Arief *et al.*, 2015). In a *p*-rep design for a 2-year trial with two replications in each year, 50% of the genotypes could be replicated twice in the first year and the remaining 50% of genotypes replicated twice in the second year. This design, simulated using the variance components from one of

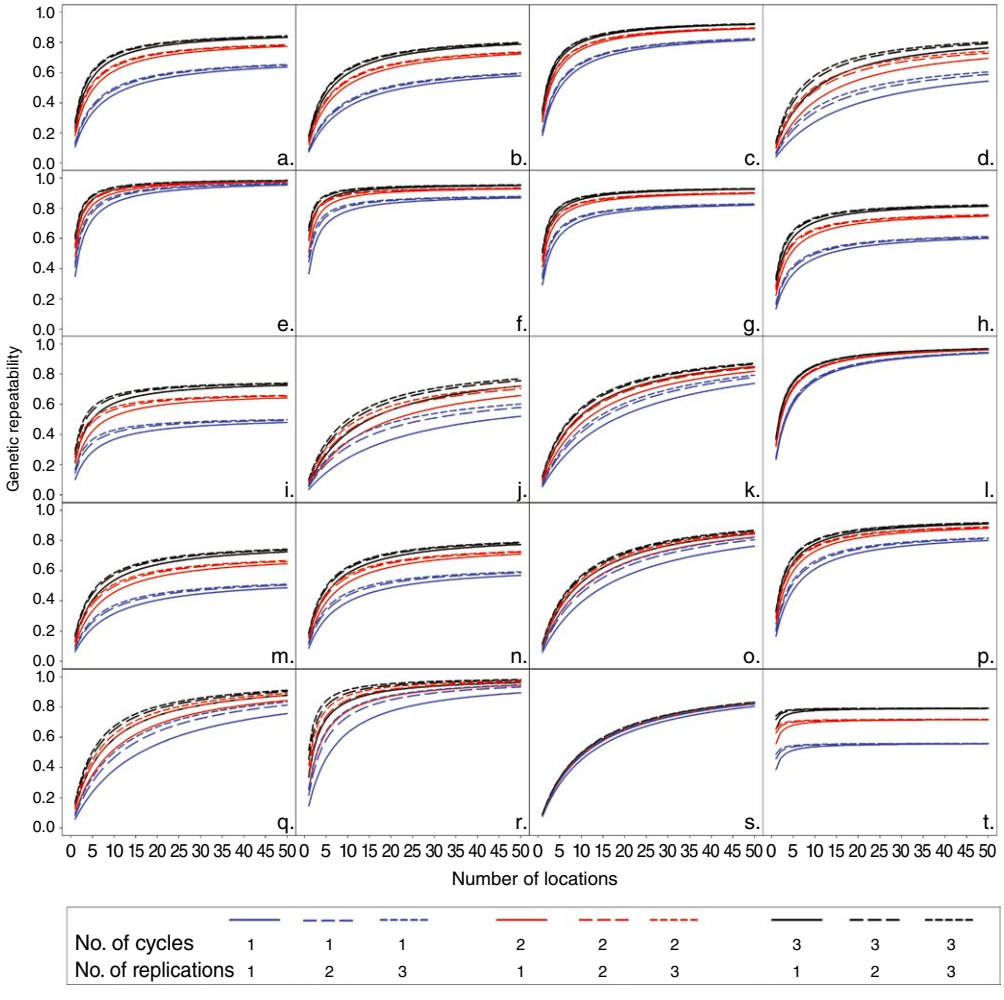


Fig. 11.2. Genetic repeatability for 20 traits modelled across a range of 1 to 50 locations, 1 to 3 replications and 1 to 3 years. Genetic repeatability was modelled using the estimated variance components for the 20 traits, as obtained from analyses of data from the first 25 cycles of Elite Spring Wheat Yield Trial (ESWYT): (a) stem rust; (b) leaf rust; (c) stripe rust; (d) grain yield; (e) 1000-kernel weight; (f) plant height; (g) days to heading; (h) test weight; (i) grain protein; (j) lodging; (k) shattering; (l) stripe rust on the spike; (m) *Septoria tritici* blotch; (n) *Septoria nodurum* blotch; (o) spot blotch; (p) powdery mildew; (q) barley yellow dwarf virus; (r) *Fusarium nivara*; (s) tan spot; and (t) black chaff. Refer to Arief *et al*. (2015) for details.

CIMMYT's international nurseries, produced better trialling efficiency than the current 1-year trial (Arief *et al*., 2015).

Analysis of Field Trials

Two levels of analyses can be employed for the field trials: (i) within-field analysis and (ii) across-fields analysis. The within-field analysis estimates genotype performance for a field, whereas the across-fields analysis estimates average genotype performance across fields, either within a year or across multiple years. Across-fields analysis can also be used to study the effect of GEI in a plant breeding programme.

Field trials are conducted for each phase of selection in a plant breeding programme for

annual crops, but the analysis is often conducted for each phase for each breeding cycle (i.e. single-set analysis). However, as genotypes tested in the same field are usually grown under the same management, the analysis should be conducted using all data from that field (i.e. single-field analysis). A single-field analysis would provide a better estimate of the field's residual variance by combining common checks across sets and would also enable a better estimate of spatial trends within the field (Arief *et al.*, 2019). Once the spatial model and residual variance are obtained for each field, combined analysis across fields could be conducted for any subsets of genotypes or fields, either as a single-year or multi-year analysis (Arief *et al.*, 2019).

In the past, a combined analysis across sets, either within or across fields (and multi-year analysis), was restricted by the imbalance in the data. The development of computers and the restricted maximum likelihood (REML, Patterson and Thompson, 1971) method for mixed-model analysis have overcome this restriction (e.g. Gilmour *et al.*, 1995; Basford *et al.*, 1996; Littell, 2002; Smith *et al.*, 2005; Piepho *et al.*, 2008).

There are two mixed-model approaches in the combined analysis across fields: a one-stage analysis and a two-stage analysis. A one-stage analysis is considered to be the 'gold standard' (Smith *et al.*, 2001b) because it provides the best linear unbiased estimators (BLUEs) of all fixed effects and the best linear unbiased predictors (BLUPs) of all random effects under the assumed single-stage model (Piepho *et al.*, 2012). In a two-stage analysis, the genotype means are estimated for each field separately in the first stage and then used (with the field residual variance and an appropriate weighting) in the second stage to estimate genotype performance across fields (Smith *et al.*, 2001a).

Single-field analysis

A single-field analysis is conducted to model the heterogeneity within a field. This modelling can be based on a spatial model or a (randomized-based) design model (Qiao *et al.*, 2004; Piepho *et al.*, 2008). Combined use of a spatial and appropriate design (e.g. row–column) model, known as the preferred model (Basford *et al.*, 1996; Gilmour *et al.*, 1997; Cullis *et al.*, 1998; Qiao *et al.*, 2004), has become common practice in the analysis of single-field data, especially for yield (whenever row and column records are available) (Smith *et al.*, 2001a; Piepho and Williams, 2010; Arief *et al.*, 2019). The preferred model has been shown to give better results than using spatial analysis alone (Cullis *et al.*, 1998; Qiao *et al.*, 2004). Cullis *et al.* (1998) observed that the REML log-likelihood increased significantly when row and column effects were taken into account in addition to spatial analysis. Qiao *et al.* (2004) observed that the preferred model could improve efficiency by an average of 28%.

The general preferred model for single-field analysis is:

$$\text{observation} \sim \text{genotype} + \text{spatial trend} + \text{residual} \qquad \text{(Eqn 11.1)}$$

This model provides an estimate of spatial trend modelled as functions of rows and columns and within-field error variance and should provide better precision for parameter estimation (Cullis and Gleeson, 1991; Brownie *et al.*, 1993; Smith *et al.*, 2001a). This model assumes independence between genotype and spatial trend. This assumption is likely to be met when an appropriate experimental design, such as a row–column design, is used.

In a one-stage analysis approach, the genotype is fitted as a random effect and the preferred spatial model for each field is recorded and used in the combined analysis across fields. In a two-stage analysis approach, the genotype is fitted as a fixed effect to obtain the BLUE for each genotype in a field. These genotype BLUEs and the residual variance for each field are required for the second-stage analysis (Cullis *et al.*, 1996b; Smith *et al.*, 2001b). However, because of a limited number of replications, it is sometimes difficult to get a reliable estimate of trend and residual variance in a field when genotype is fitted as a fixed effect. An extra step is added to this first-stage analysis to overcome this problem (Arief *et al.*, 2019). In the first step, the genotype is fitted as a random effect to obtain estimates of trend and residual variances. This step enables all genotypes to be used to estimate the trend. In the second step, these residual variances are

fixed and genotype is fitted as a fixed effect to obtain the genotype BLUEs for the second-stage analysis (Arief *et al.*, 2019). This approach has enabled spatial analysis to be applied to a field with less than ten replicated observations (Arief *et al.*, 2019).

In a large breeding programme, automated analysis is a necessity to cope with the size of data and the number of trials. An automated analysis is also required to provide timely results in the short time between harvest and retesting or seed increases of the selected genotypes. It is also required when modelling the spatial trend in each field (Arief *et al.*, 2019). The preferred model is a customized spatial model for each field (Qiao *et al.*, 2004), but it could be difficult to automatize. Therefore, a one-model-fits-all approach is recommended. This one-model-fits-all approach performed reasonably well in capturing the spatial trends in the fields used by a commercial breeding company, except for a few isolated fields where some residual structure was still observed (Arief *et al.*, 2019). A series of heatmaps generated for each field can be used as a tool to evaluate the fit of this one-model-fits-all approach in capturing the spatial field trends (Arief *et al.*, 2019). However, when a very large number of fields are involved, the differences between a one-model-fits-all and the preferred model are likely to be negligible in the recommended combined analysis (Qiao *et al.*, 2004).

Combined analysis across fields

Once a spatial model is chosen for each field, a combined analysis across fields can be conducted using a one-stage or two-stage approach. As noted earlier, a one-stage analysis is considered to be the 'gold standard' (Smith *et al.*, 2001b) and is theoretically preferred because it provides BLUEs for all fixed effects and BLUPs for all random effects under the assumed one-stage model (Piepho *et al.*, 2012). Optimum performance of a one-stage approach for genotype prediction has been demonstrated through simulation studies by Welham *et al.* (2010). A two-stage approach is less efficient than a one-stage approach because it uses a diagonal matrix from weights to approximate the variance–covariance matrix of genotype BLUEs from the first-stage analysis (Piepho *et al.*, 2012). According to Gogel *et al.* (2018), the results of the two-stage analysis will be identical to the results of the one-stage analysis if a full variance–covariance matrix of genotype BLUEs from the first-stage analysis is known. Since this variance–covariance matrix is usually not stored, they recommended a one-stage analysis of MET data and argued against a two-stage analysis of MET data with only a few trials.

However, the weighted two-stage analysis does provide acceptable results for genotype prediction (Möhring and Piepho, 2009; Welham *et al.*, 2010). There are several weighting methods developed to recover some efficiency in a two-stage analysis, but the difference between weighting methods is relatively small (Möhring and Piepho, 2009). A two-stage approach is more practical than a one-stage approach. The former requires less computing power, can handle a large amount of data, and can conveniently fit the spatial model for each field in the first-stage analysis (e.g. Cullis *et al.*, 1996b; Frensham *et al.*, 1997; Smith *et al.*, 2001a; Piepho *et al.*, 2012). It is also easier to implement in an automated system (Arief *et al.*, 2019). In an automated process, the three main outputs from the first-stage analysis (i.e. genotype BLUE, the number of replications for each genotype, and the residual variance) are stored in a database, ready to be used as inputs in any combined second-stage analysis (Arief *et al.*, 2019). This database would be updated whenever the results from new fields are available.

Combined analyses across fields can be used for many purposes. A two-stage approach enables the analysis to be completed for any subsets of genotypes and/or environments to suit the purpose of the researcher (Arief *et al.*, 2019). The three common objectives of a combined analysis of MET data are briefly discussed.

Estimating genotype performance: a mixed-model approach

The main purpose of analysing MET data is to predict future genotype performance (e.g. Smith *et al.*, 2001a; Piepho and Möhring, 2006; Piepho *et al.*, 2008). However, the prediction of genotype performance can be complicated because of the presence of GEIs. These GEIs should

be modelled and evaluated (e.g. Cooper and DeLacy, 1994; Crossa *et al.*, 2006; de la Vega *et al.*, 2007; Smith and Cullis, 2018) to improve the prediction of genotype performance.

In the early adoption of mixed model analyses of MET data, there was a question of whether the genotype effect should be classified as fixed (BLUE) or random (BLUP) (Smith *et al.*, 2005; Piepho *et al.*, 2008). In a review paper on the analysis of crop cultivar breeding and evaluation trials, Smith *et al.* (2005) argued that for selection purposes, the use of random genotype effects is appropriate because genotypes could be regarded as a random sample from a population. Piepho *et al.* (2008), in their review paper on the use of BLUP in plant breeding, agreed with the view of genotypes as a random sample from a population. They also pointed out that a desirable feature of BLUP is the ability to borrow information from relatives by exploiting a relationship matrix among genotypes.

In a plant breeding programme, genotypes are likely to be highly related and the estimated performance of a genotype can be improved by knowing the performance of its relatives (e.g. Crossa *et al.*, 2006; Oakey *et al.*, 2006; Piepho *et al.*, 2008). This relationship matrix, known as a kinship matrix (**K**), can be calculated from the pedigree information (coefficient of parentage [COP]) or molecular marker data (genomic relationship matrix [GRM]). The GRM can be calculated using several methods (e.g. Reif *et al.*, 2005; Van Raden, 2008), but they provide the same measures of genetic merit of the population (Tier *et al.*, 2015).

This **K** matrix only captures the additive component of the genetic value and is usually referred to as an additive relationship matrix (**A** matrix). This **A** matrix has been used to estimate the additive and non-additive components in wheat (Oakey *et al.*, 2006, 2007). For the prediction of F_1 hybrids, the **K** matrix must include the additive (**A**) component and the non-additive, i.e. dominant (**D**), component. Unlike inbred crops, both the **A** and **D** matrices for F_1 hybrids must be calculated from the additive relationship among their inbred parents (e.g. Bernardo, 1996a,b). The **D** matrix between hybrid *i*, which has inbred parents *Y* and *Z* and hybrid *j*, which has inbred parents *U* and *V* is as follows (Oakey *et al.*, 2007):

$$D_{ij} = \begin{cases} 1 - 0.5A_{YZ}, & i = j \\ 0.25\left(A_{YU}A_{ZV} + A_{YV}A_{ZU}\right) & \\ \left(1 - 0.5A_{YZ}\right)\left(1 - 0.5A_{UV}\right), & i \neq j \end{cases}$$

(Eqn 11.2)

where D_{ij} is the element of the **D** matrix for hybrids *i* and *j*; A_{YZ} and A_{UV} are the additive relationship coefficients between the parents of hybrid *i* and hybrid *j*, respectively; A_{YU} and A_{ZV} are the additive relationship coefficients between the two female parents and between the two male parents, respectively; A_{YV} is the additive relationship coefficient between the female parent of hybrid *i* and male parent of hybrid *j*; and and A_{ZU} is the additive relationship coefficient between the male parent of hybrid *i* and female parent of hybrid *j*. If the female and male parents are unrelated (i.e. $A_{YZ} = A_{UV} = A_{YV} = A_{ZU} = 0$), this formula reduces to:

$$D_{ij} = \begin{cases} 1, & i = j \\ 0.25\left(A_{YU}A_{ZV}\right), & i \neq j \end{cases}$$

(Eqn 11.3)

In maize, the COP between a female inbred and a male inbred is likely to be zero, as they are from two different heterotic groups. However, their GRM is not zero, as they share some common alleles. The **K** matrix between hybrid *i*, which has inbred parents *Y* and *Z*, and hybrid *j*, which has inbred parents *U* and *V*, can be calculated by combining the **A** matrix for parents and the **D** matrix for hybrids following the below formula from Bernardo (1996b):

$$K_{ij} = A_{YU}V_{GCA(1)} + A_{ZV}V_{GCA(2)} + D_{ij}V_{SCA}$$

(Eqn 11.4)

where K_{ij} is the element of the **K** matrix for hybrids *i* and *j*; A_{YU} and A_{ZV} are the additive relationship coefficients between the two female parents and between the two male parents, respectively; D_{ij} is the element of the **D** matrix for hybrids *i* and *j*; $V_{GCA(1)}$ is the variance of the general combining ability among females; $V_{GCA(2)}$ is the variance of the general combining ability among males; and V_{SCA} is the variance of the specific combining ability.

The inclusion of a **K** matrix, either derived from pedigree or marker information enables prediction of non-tested genotypes (e.g. Bernardo,

1996b; Jarquín *et al.*, 2014; Saint Pierre *et al.*, 2016). However, since the availability of dense molecular markers, the marker-based **K** matrix (GRM) has been shown to be more predictive than the pedigree-based **K** matrix (COP) (Crossa *et al.*, 2010).

Estimating genotype performance: genotype-by-environment interaction

METs are designed to predict the performance of genotypes in a specified TPE. For this purpose, genotypes should be tested in a supposedly random sample of environments (i.e. many locations across some years). Based on sampling theory, the average performances of these genotypes are the best predictors of their future performances in the TPE; hence the genotype with the highest average performance is usually recommended for that TPE. However, if the TPE is ecologically heterogeneous and there is substantial GEI, the genotype with the best average performance is often not the best in all environments (Piepho *et al.*, 1998). In such cases, a more specific recommendation can be made by modelling and predicting the interaction (Piepho *et al.*, 1998; Smith and Cullis, 2018). Traditionally, genotypes are compared on the basis of their average performance across environments. However, it is often more useful to compare the patterns of genotype response across environments (e.g. Cooper and DeLacy, 1994; Basford and Tukey, 1999; Yan *et al.*, 2000).

There are several ways to model GEI. Piepho *et al.* (1998) used a regression-based approach to predict genotype performance using covariate information on locations, such as rainfall and soil type. A factor analytic (FA) model has been used in the mixed model context to model the GEI (e.g. Smith *et al.*, 2015; Gogel *et al.*, 2018; Smith and Cullis, 2018). It is an extension of principal component analysis (PCA) into a mixed-model approach and has been adopted by the Australian National Variety Trials system for the analysis of their MET data (Smith *et al.*, 2015; Gogel *et al.*, 2018).

An FA model uses a relationship matrix among environments calculated from the data (i.e. intrinsic relationship matrix), whereas environmental covariates, such as soil type and rainfall, can be used to calculate an extrinsic relationship matrix (**W**) (Jarquín *et al.*, 2014; Saint Pierre *et al.*, 2016). The use of either an FA model or a **W** matrix provides a prediction of genotype performance in the related environments where those genotypes were not tested. However, a **W** matrix can also be used to predict future environments based on predicted environmental covariates. Both an FA model and a **W** matrix can be used in conjunction with the **K** matrix (Jarquín *et al.*, 2014; Saint Pierre *et al.*, 2016; Smith and Cullis, 2018) to provide an ultimate prediction model: the performance of non-tested genotypes in future environments.

Estimating variance components: single-year versus multi-year analysis

Analysis of MET data can also be used to obtain reasonable estimates of the variance components. These are essential as the optimality of the BLUP is based on the assumption that the variance components in the model are known (Smith *et al.*, 2005). Therefore, better estimates of the variance components provide a better prediction of genotype performance.

In annual crops, the prediction of genotype performance is usually calculated from the analysis of data from a single-year MET. Several studies have shown that a single-year analysis often overestimates VG and VGL, as these estimates are confounded with the estimates of VGY and VGYL (e.g. Nyquist and Baker, 1991; Holland and Nyquist, 2010; Arief *et al.*, 2015, 2019). In contrast, a multi-year analysis provides estimates of VG, VGY, VGL and VGYL (e.g. DeLacy *et al.*, 2010; Barrero Farfan *et al.*, 2013; Arief *et al.*, 2015). A simulation study by Arief *et al.* (2019) also showed that the correlation between the predicted and the true genotype performances was higher for a multi-year analysis than for a single-year analysis.

Multi-year analysis is not yet the standard procedure for the analysis of MET data for annual crops. This is possibly because of highly unbalanced datasets, with a large proportion of empty cells in the genotype-by-location-by-year data arrays. In mixed-model procedures, such as REML (Patterson and Thompson, 1971), this imbalance is no longer a problem. However, there is a concern of potential bias in variance component estimates, as these empty cells are not missing at random (Piepho and Möhring, 2006). Piepho and Möhring (2006) showed that there was no bias providing that all data used in

the selection process are included in the analysis. The simulation study by Arief *et al.* (2019) also detected no bias in the variance component estimates.

The estimates of variance components from a multi-year data analysis can be used to evaluate the size of genotype-by-environment variance in a breeding programme (e.g. Cullis *et al.*, 1996b; Barrero Farfan *et al.*, 2013; Arief *et al.*, 2015); to evaluate the trialling efficiency (e.g. Sprague and Federer, 1951; Patterson *et al.*, 1977; Cullis *et al.*, 1996b; Talbot, 1997); and for simulation studies (e.g. Piepho and Möhring, 2006; Arief *et al.*, 2015, 2019).

Description, Presentation and Summarization

The use of graphical procedures, integrated with pattern analysis methods, enables detailed presentation, description and summarization of MET data (e.g. DeLacy and Cooper, 1990; Cooper and DeLacy, 1994; Basford and Cooper, 1998; de la Vega and Chapman, 2001). Pattern analysis is the combined use of classification and ordination techniques (Williams, 1976; DeLacy *et al.*, 1996). These methods are especially powerful when investigating a large amount of data resulting from the increasing large field trials conducted in modern plant breeding programmes.

Presentation of the results from spatial analysis of single fields (first-stage)

The observations from a single field can usefully be presented as a heatmap of the field indexed by rows and columns of the field layout. As indicated previously, the results from a single-field analysis can be presented by a series of heatmaps representing the partitions of the estimated effects derived from the preferred model (Eqn 11.1) applied to the observations of entry performance (Arief *et al.*, 2019). These performance heatmaps (Fig. 11.3) enable a detailed assessment of the efficiency of the preferred model in accounting for spatial variability in the single-field analysis.

Graphical presentation of combined analysis across fields

The predictions for different attributes from any second-stage analysis across fields result in a three-way, three-mode genotype-by-environment-by-attribute (G×E×A) array (Fig. 11.4). This consists of a series of predictions recorded in a two-way two-mode genotype-by-environment table for each attribute. These are termed slices of the three-way array (Kroonenberg, 2008).

As in this chapter, most discussions in the literature concentrate on the examination of the slice for yield. As discussed above, these predictions are commonly derived from general mixed-model analyses, but development of Bayesian analytical analyses are increasingly being investigated. The resultant two-way tables, either genotype-by-environment, genotype-by-attribute or environment-by-attribute, can be investigated by the same methods. Each slice represents the classical genotype-by-environment table of genotype predictions of performance in multiple environments. In these cases, unlike the classical multivariate situation, each column has compatible data for a combined analysis. Hence, row statistics have normal meanings and normalization of attribute (column) data to deal with scale or data type is not required. Three two-way summary tables can be derived by averaging across the remaining mode. For the G×E table derived by averaging across attributes, normalization is required, e.g. calculating selection indices.

Graphical representation of two-way two-mode tables

Both modes can be structured by extrinsic and/or intrinsic criteria. Year-by-location stratification, classification of locations by climate, soil or disease information are examples of extrinsic arrangements of environments. Extrinsic taxonomies of genotypes can be derived from pedigree, DNA marker, physiological or disease information. Such structures lend themselves to graphical representation. For example, the column data in a genotype-by-environment table are well summarized by boxplots, as they integrate information on

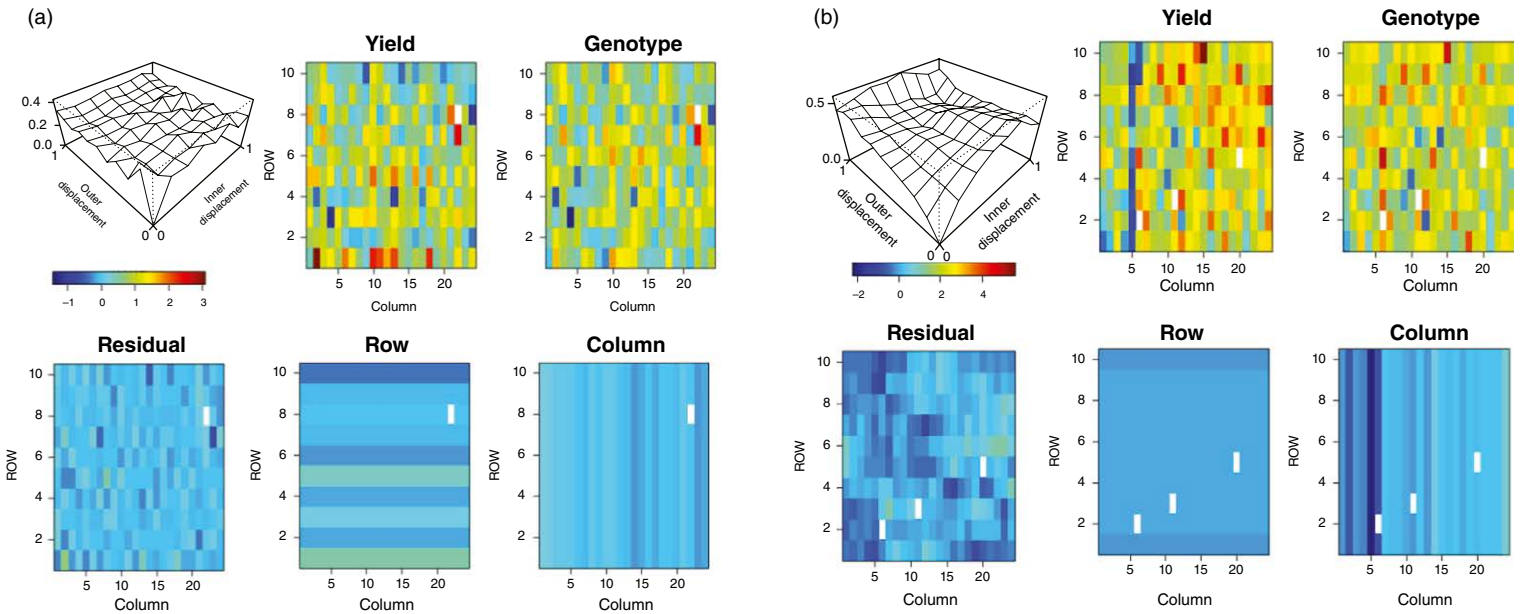


Fig. 11.3. Results from the first-stage analysis of two examples of CIMMYT Australia ICARDA Germplasm Evaluation (CAIGE) yield trial Barley 2018: (a) Narrabri and (b) Gatton. Each yield trial has six plots: a variogram to display the autoregression (*AR*1) correlation model for residual, a heatmap for centred yield data (t ha^{-1}), a heatmap for genotype effects (BLUEs) (Genotype), a heatmap for residual effects (Residual), a heatmap for row effects (Row), and a heatmap for column effects (Column). The sum of row and column effects represents the spatial trend effects (http://www.caigeproject.org.au/germplasm-evaluation/barley/yield-trial-australia/caige-yield-trial-barley-2018/).

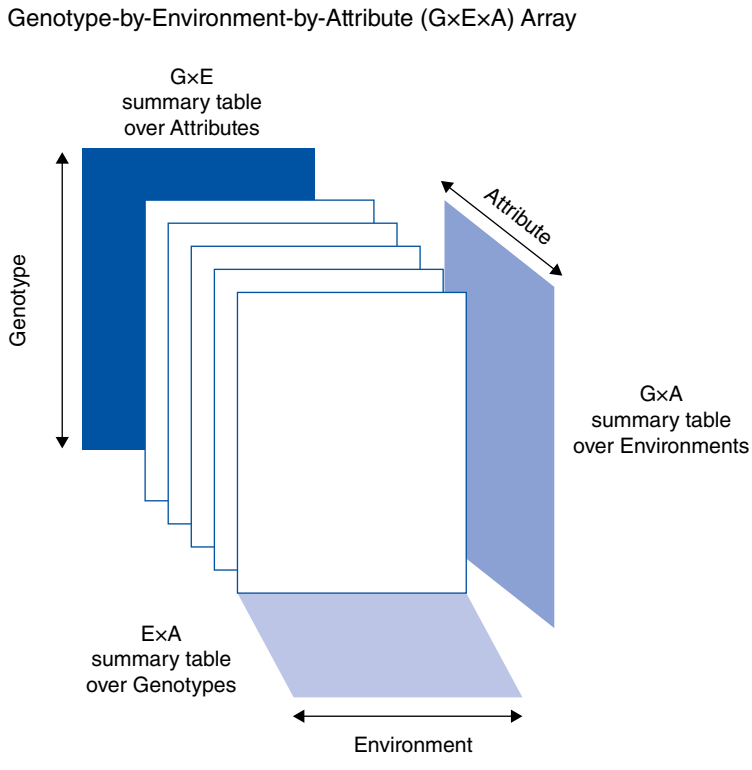


Fig. 11.4. A three-way, three-mode array of genotype-by-environment-by-attribute (G×E×A) predicted performances derived from multi-environment trials when several attributes are measured.

magnitude, variation and symmetry of the data distribution (Fig. 11.5).

Intrinsic structure in the two-way tables can also be examined by pattern analysis using an appropriate graphical representation. Most examples of classification have used an intrinsic agglomerative hierarchical classification, which is graphically displayed as a dendrogram. It is recommended that these dendrograms be optimized by a 'seriation' technique, which enables the dendrogram to be read across the base as well as the classification structure being read up and down the hierarchy (Gruvaeus and Wainer, 1972). The dendrogram is then referred to as an optimized dendrogram (Arief *et al.*, 2017). The order of the entities across the base represents a minimum distance pathway through the *n* dimensional scatter space.

A biplot (Gabriel, 1971) is used to display a low dimensional representation of the *n* dimensional scatter space representing the appropriate two-way table. It can be derived using either principal coordinate analysis using either both left and right spectral decomposition principle component analysis (through singular value decomposition). The biplots display both modes with entities (rows) displayed as points and variables (columns) displayed as vectors. The biplots can also be used to display groups of entities and groups of variables, and are then called classification-enhanced biplots. Hence, they are useful to summarize a large number of genotypes and environments in a plant breeding programme.

Graphical representation of three-way, three-mode arrays

Useful displays of a genotype-by-environment-by-attribute (G×E×A) array are achieved by appropriate investigation of the three types of summary two-way tables (Fig. 11.4). However, these analyses lose the information contained in the appropriate covariance matrices among the modes averaged across and the modes retained.

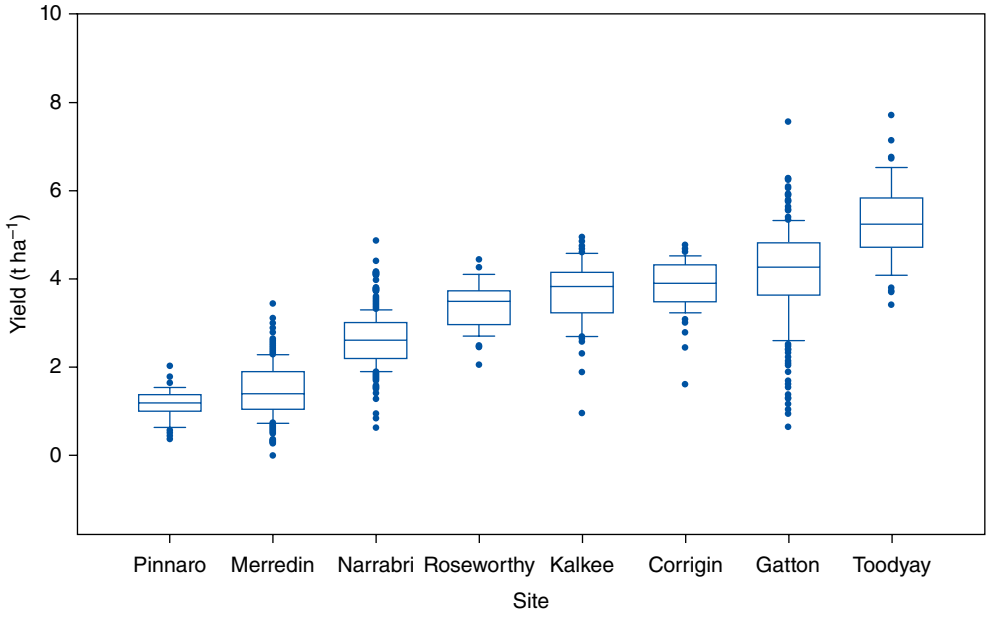


Fig. 11.5. Boxplots for grain yield (t ha^{-1}) from CIMMYT Australia ICARDA Germplasm Evaluation (CAIGE) yield trial Barley 2018. These yield trials were tested in eight sites across Australia (http://www.caigeproject.org.au/germplasm-evaluation/barley/yield-trial-australia/caige-yield-trial-barley-2018/).

Both three-way ordination and clustering techniques, together with their graphical representations, which retain this information, have been developed and proposed for the analysis of G×E×A data (Kroonenberg and Basford, 1989; Basford *et al*., 1991; Chapman *et al*., 1997).

Conclusion

METs are a major component of any plant breeding programme. Analysis of MET data is used to predict genotype performance in a TPE. Therefore, these trials should use the most appropriate experimental design to maximize their power in predicting genotype performance. It is recommended that row–column designs be adopted as standard for MET. They offer flexibility and practicality for dealing with a large number of genotypes, and the ability to estimate and eliminate spatial trend and row and column effects for the whole field, which is assumed to be under common management.

Multi-year data are accumulated in a plant breeding programme from its routine MET without any extra costs. In comparison with single-year analysis, multi-year analysis provides better prediction of genotype performance, estimates of genotype-by-year and genotype-by-year-by-location interactions, and reasonable estimates of variance components. Therefore, multi-year data analysis should become the standard procedure to analyse MET data for an annual crop.

Graphical display of the results from the analysis of multi-year multi-location trials is useful to summarize the multi-dimensional data collected from plant breeding programmes (Basford and Tukey, 1999). The interpretation of these displays can be used as an aid for understanding the plant breeding programme.

References

Arief, V.N., DeLacy, I.H., Crossa, J., Payne, T., Singh, R., *et al*. (2015) Evaluating testing strategies for plant breeding field trials: Redesigning a CIMMYT international wheat nursery to provide extra genotype connection accross cycles. *Crop Science* 55, 164–177.

Arief, V.N., DeLacy, I.H., Basford, K.E. and Dieters, M.J. (2017) Application of a dendrogram seriation algorithm to extract pattern from plant breeding data. *Euphytica* 213, 85.

Arief, V.N., Desmae, H., Hardner, C., DeLacy, I.H., Gilmour, A., *et al*. (2019) Utilization of multiyear plant breeding data to better predict genotype performance. *Crop Science* 59, 480–490.

Barrero Farfan, I.D., Murray, S.C., Labar, S. and Pietsch, D. (2013) A multi-environment trial analysis shows slight grain yield improvement in Texas commercial maize. *Field Crops Research* 149, 167–176.

Basford, K.E. and Cooper, M. (1998) Genotype X environment interactions and some considerations of their implications for wheat breeding in Australia. *Australian Journal of Agricultural Research* 49, 153–174.

Basford, K.E. and Tukey, J.W. (1999) *Graphical Analysis of Multireponse Data: Illustrated with a Plant Breeding Trial*. Chapman & Hall/CRC, Boca Raton, Florida.

Basford, K.E., Kroonenberg, P.M. and DeLacy, I.H. (1991) 3-Way methods for multiattribute genotype x environment data – an illustrated partial survey. *Field Crops Research* 27, 131–157.

Basford, K.E., Williams, E.R., Cullis, B.R. and Gilmour, A. (1996) Experimental design and analysis for variety trials. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK.

Bernardo, R. (1996a) Best linear unbiased prediction of maize single-cross performance. *Crop Science* 36, 50–56.

Bernardo, R. (1996b) Best linear unibiased prediction of the perfromance of crosses between untested maize inbreds. *Crop Science* 36, 872–876.

Brownie, C., Bowman, D.T. and Burton, J.W. (1993) Estimating spatial variation in analysis of data from yield trials: A comparison of methods. *Agronomy Journal* 85, 1244–1253.

Bull, J.K., Basford, K.E., DeLacy, I.H. and Cooper, M. (1992) Classifying genotypic data from plant breeding trials: A preliminary investigation using repeated checks. *Theoretical and Applied Genetics* 85, 461–469.

Chapman, S.C., Crossa, J., Basford, K.E. and Kroonenberg, P.M. (1997) Genotype by environment effects and selection for drought tolerance in tropical maize. 2. Three-mode pattern analysis. *Euphytica* 95, 11–20.

Cochran, W.G. and Cox, G.M. (1957) *Experimental designs*. Wiley, New York.

Coombes, N.E. (2009) DiGGer: Design search tool in R. Available at: http://nswdpibiom.org/austatgen/software/ (accessed 24 October 2019).

Cooper, M. and DeLacy, I.H. (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics* 88, 561–572.

Crossa, J., Burgueño, J., Cornelius, P.L., McLaren, G., Trethowan, R., *et al*. (2006) Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science* 46, 1722.

Crossa, J., de los Campos, G., Gianola, D., Burgueno, J., Araus, J.L., *et al*. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724.

Cullis, B.R. and Gleeson, A.C. (1991) Spatial analysis of field experiments – an extension to two dimensions. *Biometrics* 47, 1449–1460.

Cullis, B.R., Thomson, F.M., Fisher, J.A., Gilmour, A.R. and Thompson, R. (1996a) The analysis of the NSW wheat variety database. I. Modelling trial error variance. *Theoretical and Applied Genetics* 92, 21–27.

Cullis, B.R., Thomson, F.M., Fisher, J.A., Gilmour, A.R. and Thompson, R. (1996b) Analysis of the NSW wheat variety database. II. Variance component estimation. *Theoretical and Applied Genetics* 92, 28–39.

Cullis, B., Gogel, B., Verbyla, A. and Thompson, R. (1998) Spatial analysis of multi-environment early generation variety trials. *Biometrics* 54, 1–18.

Cullis, B.R., Smith, A.B. and Coombes, N.E. (2006) On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* 11, 381–393.

de la Vega, A.J. and Chapman, S.C. (2001) Genotype by environment interaction and indirect selection for yield in sunflower II. Three-mode principal component analysis of oil and biomass yield across environments in Argentina. *Field Crops Research* 72, 39–50.

de la Vega, A.J., DeLacy, I.H. and Chapman, S.C. (2007) Progress over 20 years of sunflower breeding in central Argentina. *Field Crops Research* 100, 61–72.

DeLacy, I.H. and Cooper, M. (1990) Pattern analysis for the analysis of regional variety trials. In: Kang, M.S. (ed.) *Genotype-by-Environment Interaction and Plant Breeding*. Louisiana State University, Baton Rouge, Louisiana, pp. 301–334.

DeLacy, I.H., Basford, K.E., Cooper, M., Bull, J.K. and McLaren, C.G. (1996) Analysis of multi-environment trials – an historical perspective. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK.

DeLacy, I.H., Kaul, S., Rana, B.S. and Cooper, M. (2010) Genotypic variation for grain and stover yield of dryland (rabi) sorghum in India 1. Magnitude of genotype X environment interactions. *Field Crops Research* 118, 228–235.

Federer, W.T. (2005) Augmented split block experiment design. *Agronomy Journal* 97, 578–586.

Fehr, W.R. (ed.) (1987) *Principles of Cultivars Development*. Vol. 1. Macmillan, New York.

Frensham, A., Cullis, B. and Verbyla, A. (1997) Genotype by environment variance heterogeneity in a two-stage analysis. *Biometrics* 53, 1373–1383.

Frensham, A.B., Barr, A.R., Cullis, B.R. and Pelham, S.D. (1999) A mixed model analysis of 10 years of oat evaluation data: Use of agronomic information to explain genotype by environment interaction. *Euphytica* 99, 43–56.

Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.

Gilmour, A.R., Thompson, R. and Cullis, B.R. (1995) Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450.

Gilmour, A., Cullis, B.R. and Verbyla, A.P. (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* 2, 269–293.

Gogel, B., Smith, A. and Cullis, B. (2018) Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. *Euphytica* 214, 44.

Gruvaeus, G. and Wainer, H. (1972) Two additions to hierarchical cluster analysis. *British Journal of Mathematical and Statistical Psychology* 25, 200–206.

Holland, J.B. and Nyquist, W.E. (2010) Estimating and interpreting heritability for plant breeding: An update. In: Janick, J. (ed.) *Plant Breeding Reviews*. Wiley, Oxford, UK.

Jahufer, M.Z.Z. and Luo, D. (2018) DeltaGen: A comprehensive decision support tool for plant breeders. *Crop Science* 58, 1118–1131.

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P.D., Daucourt, J., *et al*. (2014) A reaction norm model for genomic selection using high-dimentional genomic and environmental data. *Theoretical and Applied Genetics* 127, 595–607.

John, J.A. and Williams, E.R. (1995) *Cyclic and Computer Generated Designs*. Chapman & Hall/CRC, Boca Raton, Florida.

Kempton, R.A. (1984) The design and analysis of unreplicated field trials. *Vorträge fur Pflanzenzüchtung* 7, 219–242.

Kempton, R.A., Seraphin, J.C. and Sword, A.M. (1994) Statistical analysis of two-dimensional variation in variety yield trials. *The Journal of Agricultural Science* 122, 335–342.

Kroonenberg, P.M. (2008) *Applied Multiway Data Analysis*. Wiley, Hoboken, New Jersey.

Kroonenberg, P.M. and Basford, K.E. (1989) An investigation of multi-attribute genotype response across environments using three-mode principal component analysis. *Euphytica* 44, 109–123.

Littell, R.C. (2002) Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics* 7, 472–490.

Möhring, J. and Piepho, H.-P. (2009) Comparison of weighting in two-stage analysis in plant breeding trials. *Crop Science* 49, 1977–1988.

Nyquist, W.E. and Baker, R.J. (1991) Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences* 10, 235–322.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. and Kuchel, H. (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics* 113, 809–819.

Oakey, H., Verbyla, A.P., Cullis, B.R., Wei, X. and Pitchford, W.S. (2007) Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics* 114, 1319–1332.

Patterson, H.D. and Robinson, D.L. (1989) Row-and-column designs with two replicates. *The Journal of Agricultural Science* 112, 73–77.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554.

Patterson, H.D. and Williams, E.R. (1976) A new class of resolvable incomplete block designs. *Biometrika* 63, 83–92.

Patterson, H.D., Silvey, V., Talbot, M. and Weatherup, S.T.C. (1977) Variability of yields of cereal varieties in U.K. trials. *Journal of Agricultural Sciences* 89, 239–245.

Piepho, H.P. and Möhring, J. (2006) Selection in cultivar trials – is it ignorable? *Crop Science* 146, 192–201.
Piepho, H.P. and Williams, E.R. (2010) Linear variance models for plant breeding trials. *Plant Breeding* 129, 1–8.
Piepho, H.-P., Denis, J.-B. and Van Eeuwijk, F.A. (1998) Predicting cultivar differences using covariates. *Journal of Agricultural, Biological, and Environmental Statistics* 3, 151–162.
Piepho, H.P., Büchse, A. and Truberg, B. (2006) On the use of multiple lattice designs and α-designs in plant breeding trials. *Plant Breeding* 125, 523–528.
Piepho, H.P., Möhring, J., Melchinger, A.E. and Buchse, A. (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228.
Piepho, H.-P., Möhring, J., Schulz-Streeck, T. and Ogutu, J.O. (2012) A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal* 54, 844–860.
Qiao, C.G., Basford, K.E., DeLacy, I.H. and Cooper, M. (2004) Advantage of single-trial models for response to selection in wheat breeding multi-environment trials. *Theoretical and Applied Genetics* 108, 1256–1264.
Reif, J.C., Melchinger, A.E. and Frisch, M. (2005) Genetical and mathematical properties of simmilarity and dissimilarity of coefficients applied in plant breeding and seed bank management. *Crop Science* 45, 1–7.
Saint Pierre, C., Burgueño, J., Crossa, J., Dávila, G.F., López, P.F., *et al*. (2016) Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. *Scientific Reports* 6, 27312.
Smith, A.B. and Cullis, B.R. (2018) Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214, 143.
Smith, A., Cullis, B. and Gilmour, A. (2001a) The analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics* 43, 129–145.
Smith, A., Cullis, B. and Thompson, R. (2001b) Analysis variety by environments data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147.
Smith, A.B., Cullis, B.R. and Thompson, R. (2005) The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *Journal of Agricultural Science* 143, 449–462.
Smith, A.B., Ganesalingam, A., Kuchel, H. and Cullis, B.R. (2015) Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics* 128, 55–72.
Sprague, G.F. and Federer, W.T. (1951) A comparison of variance components in corn yield trials: II. Error, year × variety, location × variety and variety components. *Agronomy Journal* 43, 535–541.
Talbot, M. (1997) Resources allocation for selection systems. In: Kempton, R.A. and Fox, P.N. (eds) *Statistical Methods for Plant Variety Evaluation*. Chapman & Hall, London.
Tier, B., Meyer, K. and Ferdosi, M.H. (2015) Which genomic relationship matrix? *Proceedings of Association for the Advancement of Animal Breeding and Genetics* 21, 461–464.
Van Raden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91, 4414–4423.
Welham, S.J., Gogel, B.J., Smith, A.B., Thompson, R. and Cullis, B.R. (2010) A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics* 52, 125–149.
Williams, E.R. and John, J.A. (1996) Row-column factorial designs for use in agricultural field trials. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45, 39–46.
Williams, E.R., John, J.A. and Whitaker, D. (2006) Construction of resolvable spatial row-column designs. *Biometrics* 62, 103–108.
Williams, W.T. (1976) *Pattern Analysis in Agricultural Science*. Elsevier, Amsterdam, The Netherlands.
Yan, W., Hunt, L.A., Sheng, Q. and Szlavnics, Z. (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science* 40, 597–605.
Yates, F. (1937) A further note on an arrangement of variety trials: Quasi-Latin squares. *Annals of Eugenics* 7, 319–332.
Yau, S.K. (1997) Efficiency of alpha-lattice designs in international variety yield trials of barley and wheat. *Journal of Agricultural Science* 128, 5–9.

12 Advances in the Definition of Adaptation Strategies and Yield-stability Targets in Plant Breeding

Paolo Annicchiarico*
Research Centre for Animal Production and Aquaculture, Lodi, Italy

Introduction

This chapter focuses on the analysis of genotype–environment interaction (GEI) based on available crop yield data aimed to define adaptation strategies and yield-stability targets for a region targeted by a breeding programme. The analysis of molecular data will not be considered, while assuming that the adoption of genome-based selection will still require to cope with GEI, because a substantial portion of useful markers and quantitative trait loci (QTL) could be environment-specific in the presence of large GEI for yield (Crossa, 2012). Hereafter, 'genotype' indicates a cultivar (either genetically homogeneous, such as a pure line, or heterogeneous, such as an open-pollinated population) rather than an individual's genetic make-up. 'Environment' pools the set of climatic, soil, biotic and management conditions for the crop in a given location–year (annuals) or location–crop cycle (perennials) combination.

Increasing knowledge on plant adaptation mechanisms confirms the impossibility to breed a cultivar capable of maximizing the crop yield potential across all environments of a target region. For example, alfalfa adaptation to drought-prone and to moisture-favourable environments is largely shaped by mutually incompatible traits (Annicchiarico *et al*., 2013). A genetically based trade-off between yield potential and tolerance to drought was observed in wheat and other cereals (Ludlow and Muchow, 1990), and even wheat adaptation to different patterns of drought stress depends on partly specific traits (van Ginkel *et al*., 1998). Also, different optima of phenological development may be required across target environments, to optimally match the local growing season (Wallace *et al*., 1993) or to escape locally prevailing abiotic stresses such as low winter temperatures and terminal drought (Annicchiarico and Iannucci, 2008). Sufficient knowledge on GEI patterns and relevant environmental factors and adaptive traits can help devise a breeding strategy aimed to maximize the crop yield potential under specific or prevailing cropping conditions and to minimize the occurrence of very low yields or marked inconsistency of performance across target environments. Decisions on adaptation and yield-stability targets, genetic resources, variety type and selection procedures may represent the components of this strategy (Annicchiarico, 2002, 2009).

Adaptation and Yield Stability

In evolutionary biology, adaptation is a process, adaptedness is the level of genotype adaptation

* Email: paolo.annicchiarico@crea.gov.it

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

to a given environment, and adaptability is the ability to show good adaptedness in a range of environments. In plant breeding, the first two terms relate to a condition rather than a process, indicating the genotype's ability to be high yielding in a given environment. Genotype adaptation is usually assessed according to yield responses and undergoes modification when better-performing material becomes available. Breeding for wide adaptation and for high yield stability and reliability have sometimes been considered synonymous. However, only the adaptive responses to locations, geographic areas, farming practices or other factors that can be controlled or predicted prior to sowing can be exploited by selecting and growing specifically adapted genotypes. Accordingly, some authors (e.g. Lin and Binns, 1988) proposed to apply the yield-stability concept only to genotype responses over time, using the adaptation concept for responses in space. This view, accepted here, agrees with the farmer's view that yield consistency across time is the only relevant component of a genotype's yield stability. Breeding for wide adaptation aims to develop a variety that performs well in nearly all the target region, whereas breeding for specific adaptation aims to produce different varieties, each of which performs well in a definite subregion (alias mega-environment) or crop management within the region.

Actually, only the genotype–location (GL) interaction that is repeatable in time may be exploited by selecting and growing specifically-adapted material. The non-repeatable GL interaction is the genotype–location–year (GLY) interaction in the following analysis of variance (ANOVA) model holding year and location as crossed factors (where it is the error term for testing GL effects under the usual assumption of year as a random factor):

$$R_{ijkr} = m + G_i + L_j + Y_k + GL_{ij} + GY_{ik} + LY_{jk} + GLY_{ijk} + e_{ijkr} \quad \text{(Eqn 12.1)}$$

where R_{ijkr} is the yield of the genotype i at the location j, year k and plot r, m is the grand mean, G_i, L_j and Y_k are genotype, location and year main effects, and e_{ijkr} is the random error. Genotype–year (GY) and GLY interactions may be pooled in a GY interaction within location term that acts as the error for GL interaction in the following ANOVA model holding the year factor nested into location, which is useful when locations differ for test years:

$$R_{ijkr} = m + G_i + L_j + Y_k(L_j) + GL_{ij} + GY_{ik}(L_j) + e_{ijkr}. \quad \text{(Eqn 12.2)}$$

Early plant breeders advocated the usefulness of selection for specific adaptation (e.g. Engledow, 1925), but modern plant breeding has emphasized the selection of widely adapted material, even in less developed countries, where it tended to promote varieties with high yield potential along with technological packages designed to improve the environment. This wide-adaptation prospect has been increasingly challenged. When economically convenient, breeding for specific adaptation can contribute to more sustainable agriculture by fitting cultivars to an environment (instead of altering the environment via costly and/or environment-unfriendly inputs), increasing the biodiversity of cultivated material and meeting more closely the farmers' needs (Bramel-Cox *et al.*, 1991; Ceccarelli, 1996). Crossing and hybridization operations could remain centralized at a single station that provides each subregion with material for local selection. Subregions may be identified not only within large and/or transnational regions but also within relatively small regions, as suggested by results for barley in Syria (Ceccarelli, 1996; Ceccarelli *et al.*, 1998), bread wheat (Annicchiarico and Perenzin, 1994) and faba bean (Annicchiarico and Iannucci, 2008) in Italy, alfalfa in northern Italy (Annicchiarico, 1992; Annicchiarico and Piano, 2005), bread wheat in New South Wales (Basford and Cooper, 1998) and Ontario (Yan *et al.*, 2000), durum wheat in Algeria (Annicchiarico *et al.*, 2005), and common bean in south-western Canada (Saindon and Schaalje, 1993). A specific-adaptation strategy can easily incorporate, and is reinforced by, farmer-participatory selection (Ceccarelli, 1996, 2015). Besides, it is the ideal context for evolutionary selection schemes, by which segregating material of an inbred species, or a genetically heterogeneous population of an outbred species, undergoes natural selection under conditions representing those of a target environment or crop management (e.g. Murphy *et al.*, 2005).

Repeatable GL interaction can be either exploited by breeding specifically adapted germplasm, or minimized by breeding widely adapted material. Also, the remaining GEI terms can be either exploited by breeding material that tends to maintain its yield constant across environments

(i.e. responding relatively better in unfavourable years), or minimized by breeding genotypes with no marked deviation from their expected mean in each environment (i.e. displaying minimal GEI). These contrasting features relate to two yield stability concepts referred to, respectively, as static and dynamic by Becker and Léon (1988), and as Type 1 and Type 2 by Lin *et al.* (1986). For example, the environmental variance (S^2) measures the Type 1 stability across e environments as:

$$S_i^2 = \sum(R_{ij} - m_i)^2/(e-1) \qquad \text{(Eqn 12.3)}$$

where R_{ij} is the yield in environment j, and m_i is the mean yield across environments, of genotype i. The environmental variance focuses on all GEI effects, which may be relevant when selecting for higher yield stability within a region or a large subregion. The environmental variance applied to genotype relative yields becomes a Type 2 stability measure (with relative yields affecting the genotype performance across environments similarly to a logarithmic transformation, as frequently convenient when site mean yields vary largely and include very low yields; Annicchiarico, 2002).

Lin and Binns (1988) proposed a Type 4 stability measure that belongs to the static concept but considers yield stability exclusively in time (across years or crop cycles within locations) rather than indefinite environments. Since their measure is inflated by experimental error variance, Annicchiarico (2002, 2009) proposed an unbiased measure of temporal stability variance.

There are several Type 2 stability measures. Shukla's stability variance and Wricke's ecovalence are equivalent for genotype ranking (Becker and Léon, 1988). Another measure is based on the genotype distance from the origin of statistically significant GEI principal component (PC) axes in an Additive Main effects and Multiplicative Interaction (AMMI) analysis, computed as the unsquared Euclidean distance (Annicchiarico, 1997b) or the sum of the absolute values of the genotype PC scores (Sneller *et al.*, 1997). The AMMI modelling aims to discard GEI noise due to experimental error from the assessment of yield stability, to improve its repeatability (Sneller *et al.*, 1997). Finlay and Wilkinson's (1963) regression of genotype yield as a function of the environment mean yield, if accounting for large GEI variation, could also be used as a stability measure with reduced GEI noise. The regression slope may act as a Type 1 or Type 2 stability measure by assuming zero or one, respectively, for greatest stability.

Static stability measures (of Type 1 or Type 4) may offer various advantages over dynamic measures, such as (i) somewhat higher repeatability or heritability; (ii) estimation independent from the set of tested genotypes (which allows for a broader generalization); (iii) less ambiguous agronomic interpretation; and (iv) greater relevance for increasing food security or agricultural income (Annicchiarico, 2002). A reliable assessment of yield stability requires at least nine to ten test environments, and probably many more (Mühleisen *et al.*, 2014), owing to high sampling error (Kang, 1998).

The practical interest in selecting simultaneously for high mean yield and yield stability led to development of various yield-reliability measures aimed to assess the genotype ability to display consistently high yield across target environments (Kang and Pham, 1991). Kataoka (1963) proposed a simple index based on the square root of the environmental variance, which estimates the lowest genotype yield that is expected for a probability P fixed according to the level of farmers' risk aversion. For example, $P = 0.95$ (lowest yield expected in 95% of cases) indicates high concern for disastrous events with little consideration for mean yield response. P may vary from 0.95 (for subsistence agriculture in unfavourable regions) to 0.70 (for modern agriculture in very favourable regions) (Eskridge, 1990). For genotype i, the index is:

$$I_i = m_i - Z_{(P)}S_i \qquad \text{(Eqn 12.4)}$$

where $Z_{(P)}$ is the percentile from the standard normal distribution for which the cumulative distribution function reaches the value P (e.g. $Z_{(P)}$ is 0.675 for $P = 0.75$, and 1.280 for $P = 0.90$). Kataoka's approach was extended to derive indexes for Type 2 stability measures (Eskridge, 1990), and applies to Type 2 stability when using relative yields (Annicchiarico, 1992).

A derivation of Kataoka's approach is the incorporation of temporal yield stability into genotype yield responses modelled by AMMI or factorial regression, thereby modelling yield reliability responses across sites (Annicchiarico, 2002, 2009).

An Analytical Flow Chart

The main steps to define an adaptation strategy and yield-stability targets from multi-site, multi-year yield responses of a set of genotypes are summarized in Fig. 12.1. The aim is to generate predictions for future breeding material as represented by the test genotypes. Test sites should represent the agro-environmental variation within the target region or well-defined geographic areas and, for annual crops, should include at least 2-year data.

GEI variance components relative to lack of genetic correlation and heterogeneity of genotypic variance among environments can be estimated as described by Cooper *et al.* (1996) and reported also in Annicchiarico (2002). If the latter term (which is irrelevant for breeding) is larger than the former, it should be reduced by a data transformation as discussed in Annicchiarico (2002). This is likely to occur when site mean yields vary largely and include very low yields.

The estimation of variance components (e.g. by a REML method) for genotypic and GEI effects can justify the search for candidate subregions if the GL interaction variance is significant and moderately large (≥0.30–0.35) relative to the genotypic variance (Atlin *et al.*, 2000). Two or more candidate subregions may be identified that are large enough to be of commercial interest and lend themselves to a definition based on geography, environmental factors or farming practices. Wide- and specific-adaptation scenarios can be compared by yield gains predicted from original yield data of the same dataset or other data, or observed yield gains. Wide adaptation may be preferred owing to low GL interaction variance or to high GL interaction variance with no clear advantage of specific breeding, with different implications for the choice of selection environments (the analysis can also help locate these environments).

A yield-stability target may be justified when the overall variance accounted for by the relevant GEI components is large (say, ≥2) relative to the genotypic variance. Yield-stability targets may be subregion-specific and can be affected by other considerations (e.g. costs of additional selection environments; emphasis on food security policies).

Datasets for perennials may lack repetition in time (thereby overlooking yield-stability targets), based on results for alfalfa suggesting that the environmental variation across a 3-year crop cycle is wide enough to act as a buffer against the occurrence of non-repeatable GL effects (Annicchiarico, 1992).

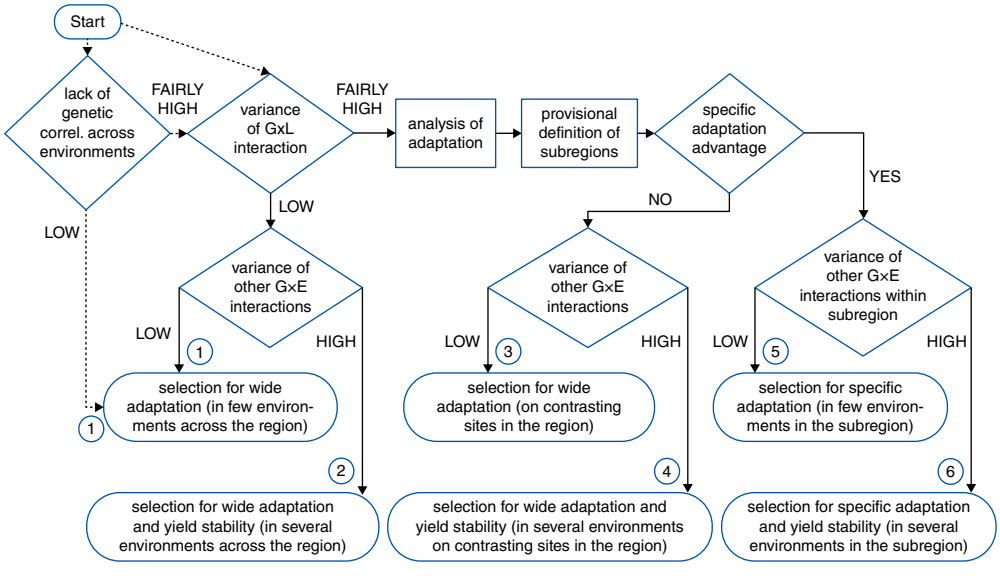


Fig. 12.1. Flow chart of steps for defining an adaptation strategy and yield-stability targets from the analysis of multi-location yield trials repeated in time (GL = genotype–location and GE = genotype–environment interactions; environment as location–year or location–crop cycle combination).

Other datasets could be relevant for defining adaptation strategies. Some may help verify to which extent GEI effects are accounted for by putative subregions with contrasting climate and/or genotypes representing contrasting plant types (e.g. Annicchiarico and Iannucci, 2008). Other datasets are suitable for comparing adaptation strategies and/or selection procedures with respect to one crop management factor (e.g. organic or conventional crop management; pure stand or intercropping; high or low nutrient availability).

Analysis of Adaptation

Techniques for the analysis of adaptation were developed with two main aims: (i) defining an adaptation strategy for breeding programmes (which may include the definition of optimal selection environments), and (ii) targeting genotypes and/or defining variety recommendations. While one dataset may serve both aims, partly different analytical approaches may be required for each aim. All GEI effects, including those of poorly performing material, are relevant to define candidate subregions for breeding and yield-stability targets. In contrast, only GL interaction effects of crossover type (i.e. implying rank change) and yield-stability differences that are relative to top-performing genotypes are relevant to define subregions for genotype targeting/recommendation.

Candidate subregions for breeding can be identified by various techniques. Pattern analysis can classify locations according to their similarity for GL interaction (DeLacy *et al.*, 1996). For trials repeated also in time, it implies a hierarchical cluster analysis performed on genotype yields averaged across time and preliminarily standardized within location. Ward's clustering method allows for site classification that reflects opportunities to exploit indirect selection among locations (Cooper *et al.*, 1996). Pattern analysis is suitable for application to largely unbalanced data sets (e.g. DeLacy *et al.*, 1994).

Modelling GL_{ij} interaction effects for yield has special interest for cultivar targeting, but can also support the definition of adaptation strategies as a step preceding site grouping or in studies comparing different germplasm types. It can be pursued by:

1. Joint regression:

$$GL_{ij} = ß_i L_j + d_{ij}$$

where $ß_i$ is Perkins and Jinks' (1968) genotype regression coefficient, equal to $(b_i - 1)$ in Finlay and Wilkinson's (1963) model; L_j is the location main effect; and d_{ij} is the residual GL interaction (whose mean square acts as the error term for heterogeneity of genotype regressions). A cut-off point between a low-yielding and a high-yielding candidate subregion for breeding may be defined by Singh *et al.*'s (1999) main crossover point, which estimates the value of site mean yield for which crossover interactions between genotypes reach the highest frequency.

2. AMMI:

$$GL_{ij} = \sum u_{in} v_{jn} l_n + d_{ij}$$

where u_{in} and v_{jn} are the eigenvectors (scaled as unit vectors) of genotypes and locations, respectively, and l_n is the square root of the eigenvalue, for $n = 1, 2, \ldots N$ axes of a double-centred principal components analysis (PCA) performed on the GL interaction matrix (Gauch, 1992). The F_R test is a simple but commendable testing criterion for PC axes (Piepho, 1995), adopting the same error term used for the overall GL interaction (Annicchiarico, 1997a). Environmental and genotypic factors associated with the occurrence of GL interaction may be revealed indirectly by correlations with GL interaction PC scores of environments and genotypes, respectively. Site classification may rely on the cluster analysis of locations as a function of their score on significant GL interaction PC axes (Annicchiarico, 1992). The Genotype main effect and GE interaction (GGE) model (Yan *et al.*, 2000) is another popular method based on singular value decomposition that may be less useful than AMMI, because it does not separate genotype main effects and GL interaction (beside tending to produce more complex graphical displays) (Gauch *et al.*, 2008).

3. Factorial regression:

$$GL_{ij} = \sum ß_{in} V_{jn} + d_{ij}$$

where $ß_{in}$ is the genotype regression on the environmental covariate n, and V_{jn} is the value on the site j of the environmental covariate, for $n = 1, 2, \ldots N$ covariates (Denis, 1988; Malosetti *et al.*, 2013). Covariates are usually quantitative, but qualitative ones can be incorporated by a set

of dummy variables (Piepho *et al.*, 1998). Regressions are usually linear, but quadratic terms may be included as additional covariates. Testing of environment covariates can rely on the error term used for the overall GL interaction (Annicchiarico, 2002). Besides environmental covariates, the model may include explicit genotypic covariates relative to traits that are either observed (Denis, 1988) or predicted genetically or by ecophysiological modelling (Malosetti *et al.*, 2013; Bustos-Korts *et al.*, 2016).

Pattern analysis is a straightforward method for site classification but does not separate pattern from noise in GL interaction effects prior to classification (where the noise relates mainly to non-repeatable GL effects). Reducing noise may be important, as shown, for example, by up to 5.6% higher yield across a target region derived from site-specific genotype targeting, and other advantages, provided by modelled data relative to observed data in an empirical assessment (Annicchiarico *et al.*, 2006).

Both efficacy (as a high proportion of GL interaction sum of squares) and parsimony (as a low number of GL interaction degrees of freedom) contribute to model value (Gauch, 1992). These features were combined into a simple criterion for model comparison equal to the sum of the estimated variances of the significant components of the GL interaction (Annicchiarico, 2002, 2009), whose model ranking agreed more closely than another simple criterion proposed by Brancourt-Hulmel *et al.* (1997) with the model's ability to predict top-yielding genotypes in an independent year (Annicchiarico, 2009). Other, more complex criteria for model comparison are also available (Malosetti *et al.*, 2013).

Cluster analysis of locations may adopt the lack of significant GL interaction within site groups as a truncation criterion (e.g. Annicchiarico, 1992), but a two-subregion scenario may be considered *a priori* as a good starting point to assess the value of breeding for specific adaptation (neglecting in any case small subregions of limited interest). An indication of the proportion of the target region occupied by each subregion is useful for comparing adaptation strategies in terms of yield gains across the target regions and other reasons (e.g. estimation of seed markets). A very rough indication can be provided by the proportion of test sites assigned to each subregion. A somewhat arbitrary upscaling of subregions may be attempted by assigning an area represented by a given test site to the subregion in which the test site was classified, or characterizing the subregions according to their mean values for environmental variables correlated with GL interaction PC scores of sites (e.g. Annicchiarico, 1992). These variables may also be exploited by a discriminant analysis of site groups aimed to provide a thorough geographical upscaling (Annicchiarico, 2002), particularly if interfaced with a geographic information system (GIS). This is shown in Fig. 12.2a for classifying sites of the potential durum wheat-growing region of Algeria into two candidate subregions according to their long-term winter mean temperature. This variable (the only significant in the discriminant analysis) accounted for 48% of the variation between site groups, and set 8°C as the truncation point for assigning sites to the cold-prone subregion A or the mild-winter subregion B (Annicchiarico *et al.*, 2002a, 2005).

For the sake of comparison, Fig. 12.2b illustrates the upscaled definition of subregions for targeting cultivars within the same Algerian region, based on factorial regression modelling of GL interaction effects as a function of site long-term values in a GIS of two statistically significant climatic covariates (Annicchiarico *et al.*, 2002b, 2006). The ability of this approach to predict top-yielding genotypes in an independent year was verified in Annicchiarico *et al.* (2006). Site-specific yield predictions for non-test environments of known genotypes (as here) or not-known ones (predicted as a function of genotype information on relevant adaptive traits or QTL) can be important for breeding programmes that cope with GL interactions *a posteriori* by targeting elite material to specific areas or cropping conditions (Bustos-Korts *et al.*, 2016).

Factorial regression has outstanding interest for fine-tuned genotype targeting, but its usefulness for defining candidate subregions for breeding may vary. For example, it proved useful in Annicchiarico *et al.* (2011) for comparing subsp. *hispanica* versus subsp. *glomerata* cocksfoot plant types across the western Mediterranean basin, where it indicated spring–summer drought stress as the essential determinant of cocksfoot GL interaction. Genotype nominal yield responses (i.e. expected yields after eliminating the site main effect) suggested two subregions for breeding, one including sites of inland

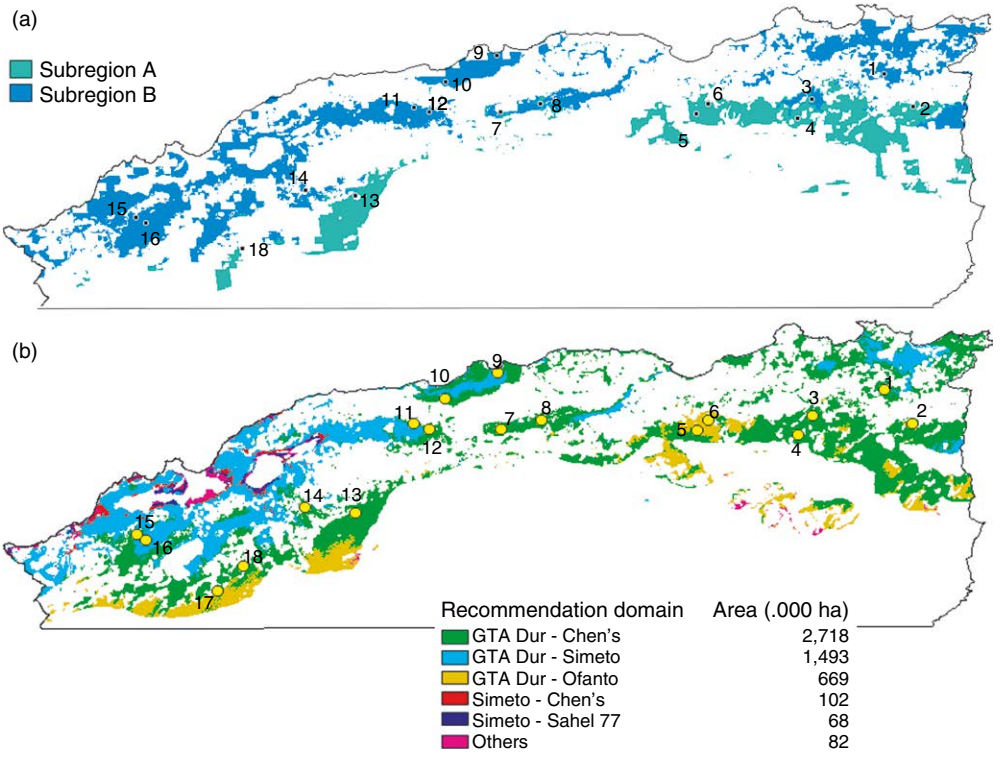


Fig. 12.2. Definition of (a) two Algerian subregions for durum wheat breeding that extends the site classification by additive main effects and multiplicative interaction + cluster analysis through the application of a discriminant function based on long-term winter mean temperature in a geographic information system (GIS); and (b) recommendation domains represented by pairs of top-yielding cultivars predicted by interfacing factorial regression modelling with long-term rainfall from October to June and winter mean temperature in a GIS. Fig. 1 and 2 reprinted, respectively, from Annicchiarico *et al*. (2002a) and Annicchiarico *et al*. (2002b).

north-west Africa that should exploit summer dormant, *hispanica* germplasm (represented by 'Kasbah'), and the other comprising sites from southern Europe and the coastal area of north-west Africa that should rely on Mediterranean *glomerata* germplasm (Fig. 12.3a). However, subregion definition based on cluster analysis of sites according to significant environmental covariates in a factorial regression analysis (with covariates possibly weighted in proportion to their importance for describing GL interaction effects) was somewhat less useful than pattern analysis and AMMI + cluster analysis for identifying Algerian subregions that could maximize the predicted advantage of durum wheat breeding for specific adaptation (Annicchiarico, 2002).

Crop growth models may contribute to define candidate subregions according to environmental factors that are the main determinants of GL interaction, as shown by the classification of the Australian wheat-growing sites according to their drought stress pattern (Chenu *et al*., 2013).

Comparison of Wide- versus Specific-adaptation Strategies

The comparison may rely on predicted or actual yield gains and varies depending on the crop breeding system and the selection procedures. A fair comparison ought to envisage similar selection costs by assuming the same total number of selection environments (as no. sites × no. years), assigning sites to subregions roughly in proportion to their relative size. Specific selection relies on entry mean yields across

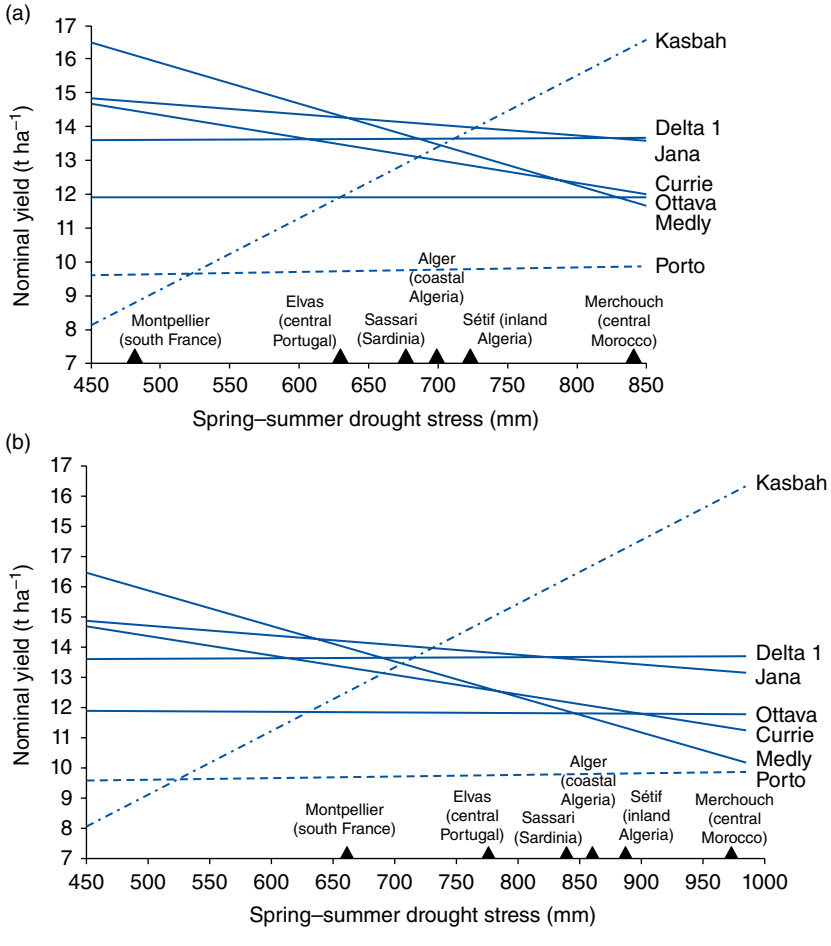


Fig. 12.3. Factorial regression-modelled nominal yield of seven cocksfoot cultivars as a function of spring–summer (April–September) drought stress of six test sites (a) on test years and (b) as predicted for 2050 according to the SRES scenario A2. Dotdash line = Mediterranean summer dormant subsp. *hispanica*; solid line = Mediterranean non-dormant or incompletely dormant subsp. *glomerata*; dotted line = Continental subsp. *glomerata*; stress as difference between estimated long-term potential evapotranspiration and actual water available for the crop. (Fig. 12.3 (a) reprinted from Annicchiarico *et al*. (2011), with permission from Elsevier.)

environments of the relevant subregion and selection for wide adaptation on entry mean yields across all environments hypothesized for specific breeding, consistent with Lin and Butler's (1988) suggestion to choose selection sites across the region in a stratified manner and in proportion to the relative size of site groups. The yield gain obtained by breeding for specific adaptation is not necessarily greater, because the advantage of exploiting positive genotype–subregion interaction effects may be offset by the disadvantage of lower precision in the estimation of entry values in each subregion due to lower number of selection environments (Atlin *et al*., 2000).

The following procedure exploits available yield data to compare adaptation strategies for inbred lines or clones in terms of predicted yield gains from one selection cycle performed in undefined selection sites. For two subregions A and B, the average predicted gain per unit area across the region provided by breeding for specific adaptation (ΔG_S) is a weighted mean of the gains ΔG_A and ΔG_B predicted for each subregion:

$$\Delta G_A = i H_A^2 s_{p(A)};\ \Delta G_B = i H_B^2 s_{p(B)};$$
$$\Delta G_S = [(\Delta G_A P_A) + (\Delta G_B P_B)] / (P_A + P_B) = (\Delta G_A P_A) + (\Delta G_B P_B) \quad \text{(Eqn 12.5)}$$

where P_A and P_B are proportions of the target region occupied by each subregion, i is the standardized selection differential, H^2 is the broad-sense heritability on a genotype mean basis estimated from genotype (s_g^2), GEI (s_{ge}^2) and pooled experimental error (s_e^2) variance components (based on data of test sites classified in the subregion) and E selection environments and R experiment replications hypothesized for selection in each subregion:

$$H^2 = s_g^2 / [s_g^2 + (s_{ge}^2/E) + (s_e^2/ER)] \quad \text{(Eqn 12.6)}$$

and s_p is the square root of the estimated phenotypic variance across environments (equal to the square root of the denominator in Eqn 12.6). The average predicted gain across the region provided by a wide-adaptation strategy is:

$$\Delta G_W = i H_{AB}^2 s_{p(AB)} \quad \text{(Eqn 12.7)}$$

where H^2 and s_p are computed from all test site data according to Eqn 12.6, setting E as the sum of the selection environments across subregions. An application of this procedure, which can easily be extended to three or more subregions, was provided by Annicchiarico *et al.* (2005) for the two Algerian subregions in Fig. 12.2a. Another procedure for the same context was proposed by Atlin *et al.* (2000). Piepho and Möhring (2005) expanded this approach by considering a more complex scenario that maximizes the selection gains by using for specific selection also the evaluation data from other subregions, giving them a weight proportional to their relevance for the target subregion.

It is also possible to compare adaptation strategies for predicted yield gains from selection in defined, nearly optimal, selection sites or in managed environments. Predicted gains are correlated gains from the defined selection environments to the target environments (DeLacy *et al.*, 1996; Basford *et al.*, 2004):

$$\Delta G_{T/S} = i r_{(S,T)} s_{p(T)} \quad \text{(Eqn 12.8)}$$

where $r_{(S,T)}$ is the phenotypic correlation for entry mean yields between selection and target environments, and $s_{p(T)}$ is the phenotypic standard deviation in the target environments. A comparison of adaptation strategies for the two Algerian subregions in Fig. 12.2a based on defined selection sites was given in Annicchiarico *et al.* (2005).

Multi-environment data of cultivar performance for outbred species bred as synthetic varieties are less useful for comparing adaptation strategies according to predicted gains, because selection mainly concerns individual plants. ΔG_W, ΔG_A and ΔG_B can be conveniently estimated for half-sib, full-sib or S_1 progeny-based selection, if multi-environment data for this material are available. For half-sib progeny testing used to select parents held as clones or selfed seed, the yield gain from one selection cycle in undefined selection environments can be predicted for each context (subregion A, subregion B and the whole region) as (Posselt, 2010):

$$\Delta G = i 0.5 s_a^2 / \sqrt{[(0.25 s_a^2) + (0.25 s_{ae}^2/E) + (s_e^2/ER)]} \quad \text{(Eqn 12.9)}$$

where s_a^2, s_{ae}^2 and s_e^2 are estimated variance components relative to the additive genetic variance, the interaction of additive genetic effects with environment and the pooled error, respectively, and E and R are hypothesized values, for the relevant context. A REML analysis performed on half-sib progeny values of relevant test environments can estimate s_e^2 and allow the estimation of the other variance components from the variance among half-sib progenies (s_g^2) and the progeny–environment interaction variance (s_{ge}^2) (assuming no inbreeding) as: $s_a^2 = 4\, s_g^2$; $s_{ae}^2 = 4\, s_{ge}^2$.

Experiment data relative to contrasting cropping conditions may contribute to define adaptation strategies. Frequently, the main interest is predicting the efficiency of indirect selection in a given condition (e.g. pure stand) relative to direct selection in a target condition (e.g. intercropping). For inbred lines or clones, such relative efficiency (E_R) can be estimated as (Falconer, 1989):

$$E_R = (i_S r_{G(S,T)} H_S) / (i_T H_T) \quad \text{(Eqn 12.10)}$$

where i_S and i_T are standardized selection differentials, and H_S and H_T are the broad-sense heritabilities on a genotype mean basis, in the selection and the target environment, and $r_{G(S,T)}$ is the genetic correlation for genotype values between the selection and target environment (which

simplifies to $E_R = r_{G(S,T)}\, H_S / H_T$, if $i_S = i_T$). For outbreds, bred as synthetic varieties, narrow-sense heritability values in the two conditions substitute for broad-sense heritability values (unless envisaging clonal evaluation). The $r_{G(S,T)}$ value can be estimated according to Robertson (1959). A predicted advantage of selection under target conditions emerged in various studies, e.g. Bänziger *et al.* (1997) for maize targeted to low-N environments.

Another avenue for comparing wide- versus specific-adaptation strategies according to predicted yield gains is by computer simulations that predict the change of allele frequencies in a few key adaptive traits across selection cycles in environments that are relevant to the different selection strategies, and its impact on yield gains. Examples were provided by Messina *et al.* (2009) and Hammer *et al.* (2014), exploiting the simulation platform QU-GENE (Podlich and Cooper, 1998).

Especially when appearing promising according to predicted yield gains, a specific-adaptation strategy may be compared with wide adaptation according to actual yield gains. For example, Ceccarelli *et al.* (1998) selected barley genotypes for wide and specific adaptation to an unfavourable (A) and a favourable (B) subregion. Actual yield gains in each subregion estimated with respect to a set of top-yielding cultivars would be $\Delta G_A = 0.03$ t ha^{-1} and $\Delta G_B = 0.08$ t ha^{-1} for specific adaptation, and $\Delta G_A = -0.03$ t ha^{-1} and $\Delta G_B = 0.08$ t ha^{-1} for wide adaptation. If the subregions had equal size, the gain per selection cycle across the region for specific (ΔG_S) and wide adaptation (ΔG_W) would be:

$$\Delta G_S = (\Delta G_A\, P_A) + (\Delta G_B\, P_B) = (0.03 \times 0.50) + (0.08 \times 0.50) = 0.055 \text{ t ha}^{-1}$$

$$\Delta G_W = (\Delta G_A P_A) + (\Delta G_B P_B) = (-0.03 \times 0.50) + (0.08 \times 0.50) = 0.025 \text{ t ha}^{-1}$$

implying 220% greater efficiency (0.055/0.025) of specific breeding. In another example relative to pure line breeding, specific selection for the two Algerian subregions in Fig. 12.2 proved over 7% more efficient than selection for wide adaptation (Annicchiarico *et al.*, 2005).

Especially for synthetic variety breeding, the lack of sufficiently large multi-environment datasets for progeny testing may give impulse to comparisons based on actual yield gains. The following example for alfalfa involved the definition and exploitation of managed selection environments. Alfalfa cultivars displayed remarkable GL interaction across sites of northern Italy that was related to soil clay content and summer water available (Annicchiarico, 1992). Three lowland geographical subregions emerged for cultivar adaptation, of which subregion A had sandy-loam soils and was moisture-favourable, subregion C had silty-clay soils and was subjected to summer drought, and subregion B had intermediate features. Four managed environments were created in one site by the factorial combination of two soil types (sandy-loam or silty-clay) by two levels of summer drought (limited or high, applied by irrigation under a moving rain-out shelter), which reproduced well the cultivar responses across agricultural sites (Annicchiarico and Piano, 2005). Wide- versus specific-adaptation to the contrasting subregions A and C were compared according to actual yield gains of Syn-2 populations selected phenotypically in the relevant managed environment(s). Results from managed environments (Annicchiarico, 2007) and agricultural sites in Table 12.1 indicated a clear advantage of specifically adapted over widely adapted selection, as only the former could provide a consistent yield progress over the locally top-performing commercial variety.

Several studies compared adaptation strategies for specific cropping conditions in terms of actual yield gains. For example, Murphy *et al.* (2007) showed an average yield advantage of 14.5% for wheat selection performed under the target condition of organic farming relative to indirect selection in conventionally managed environments.

Breeding for specific adaptation is likely to imply greater efficiency than what would emerge according to most of the described procedures, after optimization aimed to allow at least to some extent: (i) the use of a specific genetic base for each subregion, and (ii) the early allocation of novel germplasm to selection environments of only one subregion, based on observed or genomically predicted key adaptive traits. Such traits may be identified within the analysis of adaptation performed to investigate GEI for yield (e.g. Annicchiarico and Iannucci, 2008) or by further research in well-defined environments (e.g. van Ginkel *et al.*, 1998). Recent simulation models that incorporate gene action may define adaptive traits

Table 12.1. Biomass yield (t ha^{-1}) and yield gain over the locally top-performing commercial variety in three subregions as represented by managed or agricultural environments, for alfalfa phenotypic selections for wide adaptation or for specific adaptation to each of two contrasting subregions A and C. See Annicchiarico (2007) for selection details and subregion description.

	Subregion A		Subregion B[a]		Subregion C	
Selection	Yield (t ha^{-1})	Gain (%)	Yield (t ha^{-1})	Gain (%)	Yield (t ha^{-1})	Gain (%)
Managed environments[b]						
Specific for subregion A	37.59	+12.3	33.09	+22.5	–	–
Specific for subregion C	–	–	29.07	+7.7	28.38	+5.9
Widely adapted	34.24	+2.3	30.34	+12.4	28.57	+6.6
Agricultural environments[c]						
Specific for subregion A	23.26	+7.8	–	–	–	–
Specific for subregion C	–	–	–	–	34.09	+1.1
Widely adapted	19.56	–9.4	–	–	32.19	–4.5

[a]Specific selection for this intermediate subregion relies on top-performing material selected for subregions A or C.
[b]Based on data from Annicchiarico (2007). Top-performing varieties are 'Lodi' for subregion A (33.47 t ha^{-1}), and 'Prosementi' for subregions B (27.00 t ha^{-1}) and C (26.79 t ha^{-1}).
[c]Crop yield free of weeds under organic management. Top-performing varieties are 'Lodi' for subregion A (21.58 t ha^{-1}), and 'Prosementi' for subregion C (33.72 t ha^{-1}).

by predicting the impact of single traits or trait combinations on genotype adaptive responses to different subregions or contrasting environments (Chapman *et al.*, 2003; Messina *et al.*, 2009).

Definition of Selection Environments

Optimal selection environments are agricultural sites or managed environments with excellent screening ability for the target region (wide adaptation) or subregion (specific adaptation). The screening ability is proportional to the phenotypic correlation for genotype yields between selection and target environments (Eqn 12.8). Preliminary indications on useful selection environments may derive from the ordination of test sites for GL interaction effects in an analysis of adaptation.

Defining optimal selection environments is crucial when breeding for wide adaptation in the presence of large GL or GE interaction, for which three strategies were proposed. One is the definition of a few agricultural sites that contrast for GL interaction, where simultaneous or alternate selection should be carried out (also known as 'shuttle selection') (e.g. Kirigwi *et al.*, 2004). A second strategy (Podlich *et al.*, 1999) implies the classification of a large sample of target environments according to GEI effects, identifying a few major groups whose relative frequency is estimated and which are characterized by the response of some probe genotypes or a definite value of some crucial climatic variable(s). Each new selection environment is classified accordingly and is given a weight for future multi-environment selection that is proportional to the frequency of its group. A third strategy involves the definition of a set of managed selection environments capable of jointly accounting for the complexity of GEI effects or reproducing the genotype yield response under key, well-defined environmental conditions, as in Cooper *et al.* (1995), or in Annicchiarico and Piano (2005) in their wide-adaptation prospect. Federer and Scully (1993) proposed statistical designs to select material for wide adaptation across a factorial combination of two or three management or physical factors that reproduce the variation for environmental variables associated with GEI. Managed environments with controlled water availability are especially valuable in the presence of high year-to-year rainfall variation.

Methods to optimize the number of selection sites, years and experiment replications as a function of yield gains predicted from genotypic and GEI variance components are described elsewhere (e.g. Cooper *et al.*, 1999). The quality

of selection data can be enhanced by estimating best linear unbiased predictors (BLUP) entry means as a function of site-specific spatial parameters and other parameters relative to trial design, genotype, environment and GEI effects (Smith *et al.*, 2001).

Climate Change and Genotype–environment Interactions

Climate change is increasing the year-to-year climatic variability as well as the mean temperature and, in some regions, the extent of drought stress. A relevant question is whether it may prompt breeding programmes to redefine adaptation strategies and yield-stability targets, by modifying genotypic and GEI variance components. This hypothesis was verified by variance components estimated by a REML procedure for grain yield of three durum wheat cultivars grown in 11 Italian sites (spanning from northern Italy to Sicily) in all possible datasets of five consecutive years, ranging from 1991–1995 to 2007–2011 (Fig. 12.4). The ratio of GL interaction to total GEI variance averaged 0.25 until the period 1999–2003, and 0.07 in the following periods. The ratio of GL interaction to genotypic variance averaged 5.67, and usually exceeded 0.40 (a value encouraging the investigation of a specific-adaptation strategy), until 1999–2003, while averaging 0.19 and always being <0.40 in the following periods. These results imply a reduced scope for a specific-adaptation strategy, whose application to a relatively cold-prone subregion (including northern Italy and inland central Italy) and a mild-winter, Mediterranean subregion had been supported by datasets from the late 1980s or early 1990s for different cool-season cereals (Annicchiarico and Perenzin, 1994; Annicchiarico, 1997b). In contrast, greater yield stability remained a useful target, based on the mostly twofold greater temporal GEI variance relative to the genotypic variance.

Possible modifications of adaptation strategies caused by climate change may be explored also with respect to future expected values of key environmental covariates in a factorial regression model. For example, nominal yield responses of cocksfoot varieties as a function of spring–summer drought stress predicted on test sites for 2050 according to the SRES scenario A2 delivered quite different results from those relative to test years, highlighting quite a large subregion (including nearly all areas of southern Europe and north-west Africa) that should rely on the breeding of summer dormant, *hispanica* germplasm in this case (Fig. 12.3).

Conclusions and Perspectives

There are several techniques for defining adaptation strategies and yield-stability targets according to available crop yield data that serve their purposes well and are unlikely to undergo distinct progress in the next few decades. The ability of ordinary breeders and smaller-scale breeding programmes to exploit these techniques increased after the development of freely available, user-friendly software, such as PB Tools and CropStat issued by the International Rice Research Institute. Within this favourable context, however, there is insufficient availability of scientific studies showing the positive impact on crop yields of well-defined adaptation strategies or the selection for higher yield stability. This information can be important to support and motivate innovation in this direction by breeding programmes, as necessary to turn the vast body of methods and tools for GEI analysis into crops that are more productive, more resilient to prevailing stresses and more adapted to climate change. The selection for yield stability is hindered by its requirement for many selection environments, but increasing knowledge on GEI patterns can help design sets of contrasting managed selection environments that jointly reproduce and maximize the GEI effects expected across agricultural environments of the target region.

The ecophysiological modelling of different genotype–environment combinations based on crop growth models interfaced with simulation models will increasingly contribute to investigations of adaptation strategies and adaptive traits. However, its need for a suitable crop model that incorporates physiological effects and associated environmental responses for traits that influence the genetic variation for crop yield, and the phenotyping effort required to reliably parameterize the model (Cooper *et al.*, 2016), suggest that it will initially be limited to major crops.

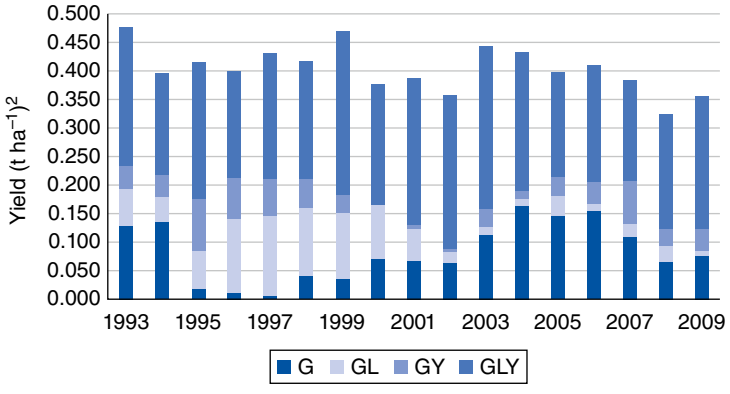


Fig. 12.4. Estimation of components of variance relative to genotype (G) and genotype–location (GL), genotype–year (GY) and genotype–location–year (GLY) interactions for grain yield of three durum wheat cultivars in 11 Italian locations in datasets of five consecutive years ranging from 1991–1995 to 2007–2011 (reporting on the *x*-axis the central year). (Based on data of the Italian network of variety trials provided by F. Quaranta and A. Belocchi.)

Breeding for specific cropping conditions will profit from high-throughput phenotyping techniques and/or genomic information able to predict key adaptive traits used as genotypic covariates in factorial regression models (Bustos-Korts *et al.*, 2016) or in crop growth models accounting for site-specific genotype differences in yield (Cooper *et al.*, 2016). Likewise, genomic selection models that incorporate GEI for yield (Crossa *et al.*, 2017) could be exploited for early allocation of putative elite material to specific environments for field-based selection. However, genomic selection may be exploited also to breed *a priori* for distinct subregions (Lado *et al.*, 2016). Genomic selection could also facilitate the selection for yield stability, according to its moderate predictive accuracy for Finlay and Wilkinson's genotype regression and AMMI analysis-derived stability (Wang *et al.*, 2015; Huang *et al.*, 2016).

Acknowledgements

I am grateful to Fabrizio Quaranta and Andreina Belocchi for providing plot yield data of the Italian network of durum wheat variety trials from test years 1991 through 2011.

References

Annicchiarico, P. (1992) Cultivar adaptation and recommendation from alfalfa trials in northern Italy. *Journal of Genetics and Breeding* 46, 269–278.

Annicchiarico, P. (1997a) Additive main effects and multiplicative interaction (AMMI) of genotype–location interaction in variety trials repeated over years. *Theoretical and Applied Genetics* 94, 1072–1077. DOI: 10.1007/s001220050517.

Annicchiarico, P. (1997b) Joint regression vs. AMMI analysis of genotype–environment interactions for cereals in Italy. *Euphytica* 94, 53–62. DOI: 10.1023/A:1002954824178.

Annicchiarico, P. (2002) *Genotype × Environment Interactions: Challenges and Opportunities for Plant Breeding and Cultivar Recommendations*. FAO Plant Production and Protection Paper No. 174. FAO, Rome, 115 pp.

Annicchiarico, P. (2007) Wide- versus specific-adaptation strategy for lucerne breeding in northern Italy. *Theoretical and Applied Genetics* 114, 647–657. DOI: 10.1007/s00122-006-0465-1.

Annicchiarico, P. (2009) Coping with and exploiting genotype × environment interactions. In: Ceccarelli, S., Guimarães, E.P. and Weltzien, E. (eds) *Plant Breeding and Farmer Participation*. FAO, Rome, pp. 519–564.

Annicchiarico, P. and Iannucci, A. (2008) Breeding strategy for faba bean in southern Europe based on cultivar responses across climatically contrasting environments. *Crop Science* 48, 983–991. DOI: 10.2135/cropsci2007.09.0501.

Annicchiarico, P. and Perenzin, M. (1994) Adaptation patterns and definition of macro-environments for selection and recommendation of common-wheat genotypes in Italy. *Plant Breeding* 113, 197–205. DOI: 10.1111/j.1439-0523.1994.tb00723.x.

Annicchiarico, P. and Piano, E. (2005) Use of artificial environments to reproduce and exploit genotype × location interaction for lucerne in northern Italy. *Theoretical and Applied Genetics* 110, 219–227. DOI: 10.1007/s00122-004-1811-9.

Annicchiarico, P., Chiari, T., Delli, G., Doucene, S., Yallaoui-Yaïci, N., *et al*. (2002a) Response of durum wheat cultivars to Algerian environments. IV. Implications on a national breeding strategy. *Journal of Agriculture and Environment for International Development* 96, 227–259.

Annicchiarico, P., Bellah, F., Chiari, T. and Delli, G. (2002b) Response of durum wheat cultivars to Algerian environments. III. GIS-based definition of cultivar recommendations. *Journal of Agriculture and Environment for International Development* 96, 209–226.

Annicchiarico, P., Bellah, F. and Chiari, T. (2005) Defining subregions and estimating benefits for a specific-adaptation strategy by breeding programs: A case study. *Crop Science* 45, 1741–1749. DOI: 10.2135/cropsci2004.0524

Annicchiarico, P., Bellah, F. and Chiari, T. (2006) Repeatable genotype × location interaction and its exploitation by conventional and GIS-based cultivar recommendation for durum wheat in Algeria. *European Journal of Agronomy* 24, 70–81. DOI: 10.1016/j.eja.2005.05.003.

Annicchiarico, P., Pecetti, L., Bouzerzour, H., Kallida, R., Khedim, A., *et al*. (2011) Adaptation of contrasting cocksfoot plant types to agricultural environments across the Mediterranean basin. *Environmental and Experimental Botany* 74, 82–89. DOI: 10.1016/j.envexpbot.2011.05.002.

Annicchiarico, P., Pecetti, L. and Tava, A. (2013) Physiological and morphological traits associated with adaptation of lucerne (*Medicago sativa* L.) to severely drought-stressed and to irrigated environments. *Annals of Applied Biology* 162, 27–40. DOI: 10.1111/j.1744-7348.2012.00576.x.

Atlin, G.N., Baker, R.J., McRae, K.B. and Lu, X. (2000) Selection response in subdivided target regions. *Crop Science* 40, 7–13. DOI: 10.2135/cropsci2000.4017.

Bänziger, M., Betrán, F.J. and Lafitte H.R. (1997) Efficiency of high-nitrogen selection environments for improving maize for low-nitrogen target environments. *Crop Science* 37, 1103–1109. DOI: 10.2135/cropsci1997.0011183X003700040012x.

Basford, K.E. and Cooper, M. (1998) Genotype × environment interactions and some considerations of their implications for wheat breeding in Australia. *Australian Journal of Agricultural Research* 49, 153–174. DOI: 10.1071/A97035.

Basford, K.E., Federer, W.T. and DeLacy, I.H. (2004) Mixed model formulation for multi-environment trials. *Agronomy Journal* 96, 143–147. DOI: 10.2134/agronj2004.0143.

Becker, H.C. and Léon, J. (1988) Stability analysis in plant breeding. *Plant Breeding* 101, 1–23. DOI:10.1111/j.1439-0523.1988.tb00261.x.

Bramel-Cox, P.J., Barker, T., Zavala-Garcia, F. and Eastin, J.D. (1991) Selection and testing environments for improved performance under reduced-input conditions. In: Sleper, D., Bramel-Cox, P.J. and Barker, T. (eds) *Plant Breeding and Sustainable Agriculture: Considerations for Objectives and Methods*. CSSA Special Publication 18, ASA, CSSA, SSSA, Madison, Wisconsin, pp. 29–56.

Brancourt-Hulmel, M., Biarnès-Dumoulin, V. and Denis, J.B. (1997) Points de repère dans l'analyse de la stabilité et de l'interaction génotype–milieu en amélioration des plants. *Agronomie* 17, 219–246. DOI: 10.1051/agro:19970403.

Bustos-Korts, D., Malosetti, M., Chapman, S. and van Eeuwijk, F. (2016) Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics. In: Yin, X. and Struik, P. (eds) *Crop Systems Biology*. Springer, Cham, Switzerland, pp. 55–82.

Ceccarelli, S. (1996) Positive interpretation of genotype by environment interaction in relation to sustainability and biodiversity. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK, pp. 467–486.

Ceccarelli, S. (2015) Efficiency of plant breeding. *Crop Science* 55, 87–97. DOI: 10.2135/cropsci2014.02.0158.

Ceccarelli, S., Grando, S. and Impiglia, A. (1998) Choice of selection strategy in breeding barley for stress environments. *Euphytica* 103, 307–318. DOI: 10.1023/A:1018647001429.

Chapman, S., Cooper, M., Podlich, D. and Hammer, G. (2003) Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agronomy Journal* 95, 99–113. DOI: 10.2134/agronj2003.0099.

Chenu, K., Deihimfard, R. and Chapman, S.C. (2013) Large-scale characterization of drought pattern: A continent-wide modelling approach applied to the Australian wheatbelt – spatial and temporal trends. *New Phytologist* 198, 801–820. DOI: 10.1111/nph.12192.

Cooper, M., Woodruff, D.R., Eisemann, R.L., Brennan, P.S. and DeLacy, I.H. (1995) A selection strategy to accommodate genotype-by-environment interaction for grain yield of wheat: Managed-environments for selection among genotypes. *Theoretical and Applied Genetics* 90, 492–502. DOI: 10.1007/BF00221995.

Cooper, M., DeLacy, I.H. and Basford, K.E. (1996) Relationships among analytical methods used to study genotypic adaptation in multi-environment trials. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK, pp. 193–224.

Cooper, M., Rajatasereekul, S., Somrith, B., Sriwisut S., Immark, S., *et al*. (1999) Rainfed lowland rice breeding strategies for northeast Thailand II. Comparison of intrastation and interstation selection. *Field Crops Research* 64, 153–176. DOI: 10.1016/S0378-4290(99)00057-X.

Cooper, M., Technow, F., Messina, C., Gho, C., Totir, L.R. (2016) Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Science* 56, 1–16. DOI:10.2135/cropsci2015.08.0512

Crossa, J. (2012) From genotype × environment interaction to gene × environment interaction. *Current Genomics* 13, 225–244. DOI:10.2174/138920212800543066.

Crossa, J., Pérez-Rodríguez. P., Cuevas, J., Montesinos-López, O., Jarquín, D., *et al*. (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science* 22, 961–975. DOI: 10.1016/j.tplants.2017.08.011.

DeLacy, I.H., Fox, P.N., Corbett, J.D., Crossa, J., Rajaram, S., *et al*. (1994) Long-term association of locations for testing spring bread wheat. *Euphytica* 72, 95–106. DOI: 10.1007/BF00023777.

DeLacy, I.H., Basford, K.E., Cooper, M., Bull, J.K. and McLaren, C.G. (1996) Analysis of multi-environment data – an historical perspective. In: Cooper, M. and Hammer, G.L. (eds) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK, pp. 39–124.

Denis, J.B. (1988) Two-way analysis using covariates. *Statistics* 19, 123–132. DOI: 10.1080/02331888808802080.

Engledow, F.L. (1925) The economic possibilities of plant breeding. In: *Proceedings of the Imperial Botanical Conference*. F.T. Brooks, London, pp. 31–40.

Eskridge, K.M. (1990) Selection of stable cultivars using a safety-first rule. *Crop Science* 30, 369–374. DOI: 10.2135/cropsci1990.0011183X003000020025x.

Falconer, D.S. (1989) *Introduction to Quantitative Genetics*, 3rd edn. Longman, New York, 438 pp.

Federer, W.T. and Scully, B.T. (1993) A parsimonious statistical design and breeding procedure for evaluating and selecting desirable characteristics over environments. *Theoretical and Applied Genetics* 86, 612–620. DOI: 10.1007/BF00838717.

Finlay, K.W. and Wilkinson, G.N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research* 14, 742–754. DOI: 10.1071/AR9630742.

Gauch, H.G. (1992) *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam, The Netherlands, 278 pp.

Gauch, H.G., Piepho, H.-P. and Annicchiarico, P. (2008) Statistical analysis of yield trials by AMMI and GGE: Further considerations. *Crop Science* 48, 866–889. DOI: 10.2135/cropsci2007.09.0513.

Hammer, G.L., McLean, G., Chapman, S., Zheng, B., Doherty, A., *et al*. (2014) Crop design for specific adaptation in variable dryland production environments. *Crop and Pasture Science* 65, 614–626. DOI: 10.1071/CP14088.

Huang, M., Cabrera, A., Hoffstetter, A., Griffey, C., van Sanford, D., *et al*. (2016) Genomic selection for wheat traits and trait stability. *Theoretical and Applied Genetics* 129, 1697–1710. DOI: 10.1007/s00122-016-2733-z.

Kang, M.S. (1998) Using genotype-by-environment interaction for crop cultivar development. *Advances in Agronomy* 62, 199–252. DOI: 10.1016/S0065-2113(08)60569-6.

Kang, M.S. and Pham, H.N. (1991) Simultaneous selection for high yielding and stable crop genotypes. *Agronomy Journal* 83, 161–165. DOI: 10.2134/agronj1991.00021962008300010037x.

Kataoka, S. (1963) A stochastic programming model. *Econometrika* 31, 181–196. DOI: 10.2307/1910956.

Kirigwi, F., van Ginkel, M., Trethowan, R., Sears, R.G., Rajaram, S., *et al*. (2004) Evaluation of selection strategies for wheat adaptation across water regimes. *Euphytica* 135, 361–371. DOI: 10.1023/B:EUPH.0000013375.66104.04.

Lado B., Barrios P.G., Quincke M., Silva P. and Gutiérrez L. (2016) Modeling genotype × environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Science* 56, 2165–2179. DOI: 10.2135/cropsci2015.04.0207.

Lin, C.S. and Binns, M.R. (1988) A method for analysing cultivar × location × year experiments: A new stability parameter. *Theoretical and Applied Genetics* 76, 425–430. DOI: 10.1007/BF00265344.

Lin, C.S. and Butler, G. (1988) A data-based approach for selecting locations for regional trials. *Canadian Journal of Plant Science* 68, 651–659. DOI: 10.4141/cjps88-078.

Lin, C.S., Binns, M.R. and Lefkovitch, L.P. (1986) Stability analysis: Where do we stand? *Crop Science* 26, 894–900. DOI: 10.2135/cropsci1986.0011183X002600050012x.

Ludlow, M.M. and Muchow, R.C. (1990) A critical evaluation of traits for improving crop yields in water-limited environments. *Advances in Agronomy* 43, 107–153. DOI: 10.1016/S0065-2113(08)60477-0.

Malosetti, M., Ribaut, J.M. and van Eeuwijk, F.A. (2013) The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology* 4, 44. DOI:10.3389/fphys.2013.00044.

Messina, C., Hammer, G., Dong, Z., Podlich, D. and Cooper, M. (2009) Modelling crop improvement in a G×E×M framework via gene-trait-phenotype relationships. *Crop Physiology* 235, 581. DOI: 10.1016/B978-0-12-374431-9.00010-4.

Mühleisen, J., Piepho, H.-P., Maurer, H.P., Zhao, Y. and Reif, J.C. (2014) Exploitation of yield stability in barley. *Theoretical and Applied Genetics* 127, 1949–1962. DOI: 10.1007/s00122-014-2351-6.

Murphy, K., Lammer, D., Lyon, S., Carter, B. and Jones, S.S. (2005) Breeding for organic and low-input farming systems: An evolutionary–participatory breeding method for inbred cereal grains. *Renewable Agriculture and Food Systems* 20, 48–55. DOI: 10.1079/RAF200486.

Murphy, K.M., Campbell, K.G., Lyon, S.R. and Jones S.S. (2007) Evidence of varietal adaptation to organic farming systems. *Field Crops Research* 102, 172–177. DOI: 10.1016/j.fcr.2007.03.011.

Perkins, J.M. and Jinks, J.L. (1968) Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity* 23, 339–356. DOI: 10.1038/hdy.1968.48.

Piepho, H.P. (1995) Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. *Theoretical and Applied Genetics* 90, 438–443. DOI: 10.1007/BF00221987.

Piepho, H.P. and Möhring, J. (2005) Best linear unbiased prediction of cultivar effects for subdivided target regions. *Crop Science* 45, 1151–1159. DOI: 10.2135/cropsci2004.0398.

Piepho, H.P., Denis, J.B. and Van Eeuwijk, F.A. (1998) Predicting cultivar differences using covariates. *Journal of Agricultural, Biological, and Environmental Statistics* 3, 151–162. DOI: 10.2307/1400648.

Podlich, D.W. and Cooper, M. (1998) QU-GENE: A simulation platform for quantitative analysis of genetic models. *Bioinformatics* 14, 632–653. DOI: 10.1093/bioinformatics/14.7.632.

Podlich, D.W., Cooper, M. and Basford, K.E. (1999) Computer simulation of a selection strategy to accommodate genotype–environment interactions in a wheat recurrent selection programme. *Plant Breeding* 118, 17–28. DOI: 10.1046/j.1439-0523.1999.118001017.x.

Posselt, U.K. (2010) Breeding methods in cross-pollinated species. In: Boller, B., Posselt, U.K. and Veronesi, F. (eds) *Fodder Crops and Amenity Grasses*. Springer, New York, pp. 39–87.

Robertson, A. (1959) The sampling variance of the genetic correlation coefficient. *Biometrics* 15, 469–485. DOI: 10.2307/2527750.

Saindon, G. and Schaalje, G.B. (1993) Evaluation of locations for testing dry bean cultivars in western Canada using statistical procedures, biological interpretation and multiple traits. *Canadian Journal of Plant Science* 73, 985–994. DOI: 10.4141/cjps93-129.

Singh, M., Ceccarelli, S. and Grando, S. (1999) Genotype × environment interaction of cross-over type: Detecting its presence and estimating the crossover point. *Theoretical and Applied Genetics* 99, 988–995. DOI: 10.1007/s001220051406.

Smith, A., Cullis, B. and Thomson, R. (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. DOI:10.1111/j.0006-341X.2001.01138.x

Sneller, C.H., Kilgore-Norquest, L. and Dombek, D. (1997). Repeatability of yield stability statistics in soybean. *Crop Science* 37, 383–390. DOI: 10.2135/cropsci1997.0011183X003700020013x.

van Ginkel, M., Calhoun, D.S., Gebeyehu, G., Miranda, A., Tian-you, C., *et al*. (1998) Plant traits related to yield of wheat in early, late, or continuous drought conditions. *Euphytica* 100, 109–121. DOI: 10.1023/A:1018364208370.

Wallace, D.H., Zobel, R.W. and Yourstone, K.S. (1993) A whole-system reconsideration of paradigms about photoperiod and temperature control of crop yield. *Theoretical and Applied Genetics* 86, 17–26. DOI: 10.1007/BF00223804.

Wang, Y., Mette, M.F., Miedaner, T., Wilde, P., Reif, J.C., *et al*. (2015) First insights into the genotype–phenotype map of phenotypic stability in rye. *Journal of Experimental Botany* 66, 3275–3284. DOI: 10.1093/jxb/erv145.

Yan, W., Hunt, L.A., Sheng, Q. and Szlavnics, Z. (2000) Cultivar evaluation and mega-environment investigation based on GGE biplot. *Crop Science* 40, 597–605. DOI: 10.2135/cropsci2000.403597x.

Section II:

Intersection of Breeding, Genetics and Genomics: Crop Examples

13 Prediction with Big Data in the Genomic and High-throughput Phenotyping Era: A Case Study with Wheat Data

Paulino Pérez-Rodríguez[1], Juan Burgueño[2], Osval A. Montesinos-López[3], Ravi P. Singh[2], Philomin Juliana[2], Suchismita Mondal[2] and José Crossa[2]*

[1]*Colegio de Postgraduados, Montecillos, Edo. de México, México;* [2]*International Maize and Wheat Improvement Center (CIMMYT), México City, México;* [3]*Universidad de Colima, Colima, México*

Introduction

The selection response is the most important breeder's equation, and factors that increase the selection intensity, along with the trait's heritability and genetic diversity, as well as factors that reduce the time needed to complete a selection cycle, will increase the overall selection response and thus increase the genetic gains relative to the target traits. Simulation and empirical results have shown that genomic selection (GS) (Meuwissen *et al.*, 2001; Bernardo and Yu, 2007; Lorenzana and Bernardo, 2009) can increase genetic gains by: (i) shortening the breeding cycle (rapid selection cycle), and/or (ii) increasing the selection intensity by performing sparse field evaluation.

In plant and animal breeding, sequencing technologies that allow using abundant and cheap molecular markers have enabled GS. Early plant breeding data have shown that, compared with pedigree and marker-assisted selection, GS significantly increases prediction accuracy for low heritability traits (de los Campos *et al.*, 2009, 2010, 2012; Crossa *et al.*, 2010, 2011, 2013, 2014; Heslot *et al.*, 2012; 2014; Pérez-Rodríguez *et al.*, 2012; Hickey *et al.*, 2012; González-Camacho *et al.*, 2012, 2016; Riedelsheimer *et al.*, 2012; Zhao *et al.*, 2012; Windhausen *et al.*, 2012; Technow *et al.*, 2013). An early review of the main genomic prediction and selection activities of CIMMYT's wheat (and maize) breeding programmes was published by Crossa *et al.* (2014); since then, breeding programmes worldwide have been studying and applying GS. At the same time, extensive research has been conducted and new statistical models for incorporating pedigree, genomic and environmental covariables (climatic and meteorological data) into statistical-genetic prediction models and methods have been generated. Models incorporating genome × environment interactions have been developed to improve accuracy when predicting individuals unobserved in test environments (locations, years or a combination of both) (Burgueño *et al.*, 2012; Jarquín *et al.*, 2014; Heslot *et al.*, 2014; López-Cruz *et al.*, 2015; Crossa *et al.*, 2016). New models for assessing the genomic prediction accuracy of categorical response variables (e.g. disease measured as ordinal rates, count data, etc.) are also being developed (Montesinos-López *et al.*, 2015a,b,c, 2016).

* Email: j.crossa@cgiar.org

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

In plant breeding, GS involves predicting breeding values that comprise the parental average (half the sum of the breeding values of both parents) plus a deviation attributable to Mendelian sampling, as noted by Crossa *et al*. (2014). Genomic selection has been applied in two different contexts; one approach focuses on predicting additive effects in early generations of a breeding programme so that a short cycle with rapid recombination is achieved. The other approach consists of predicting the genotypic values of individuals, where both additive and non-additive effects determine the final commercial value of the lines; this requires predicting a large number of lines established in a sparse multi-environment field evaluation.

Within the GS context, population structure affecting prediction accuracy was initially thought to be a factor that distorts and confounds predictions between training and testing populations (Windhausen *et al*., 2012). However, de los Campos *et al*. (2015) pointed out that natural and artificial breeding populations always have different degrees of stratification because of differences in allele frequency and in linkage disequilibrium patterns of the different families, which act as a modifier effect rather than as a confounding effect, as initially thought. Daetwyler *et al*. (2015) mentioned that what is important is controlling and accounting for spurious population structures (such as those originating from admixtures) but without affecting the relatedness between individuals within and between families. Furthermore, as clarified by de los Campos *et al*. (2010) and Janss *et al*. (2012), genomic best linear unbiased predictor (GBLUP) models account for population structures and substructures between and within families (Crossa *et al*., 2013).

Factors that determine the prediction accuracy in GS are: trait heritability, number of markers, size of the training population, the relationship between training and testing sets, and genotype × environment (G×E) interaction. Including high-density marker platforms with G×E interactions increases the predictive ability of GS models (Burgueño *et al*., 2012; Heslot *et al*., 2014; Jarquín *et al*., 2014; López-Cruz *et al*., 2015). Using dense molecular markers with pedigree information increases the prediction accuracy of unobserved phenotypes. Genomic predictions have been extensively studied in elite bread wheat germplasm (de los Campos *et al*., 2009, 2010; Crossa *et al*., 2010; Gónzalez-Camacho *et al*., 2012; Heslot *et al*., 2012; Pérez-Rodríguez *et al*., 2012, 2017; López-Cruz *et al*., 2015).

Initial research on GS prediction accuracy of the complex trait 'grain yield (GY)' in wheat comprised an intermediate-to-low number of individuals (around 300–600 wheat lines) evaluated in only a few environments and genotyped with an intermediate-to-low number of molecular markers (Crossa *et al*., 2010; Pérez-Rodríguez *et al*., 2012). However, the rapid development of new, dense and cheap marker technologies has allowed the sizes of the populations under genomic predictions to substantially increase. For example, in a recent research article, Pérez-Rodríguez *et al*. (2017) included 58,798 wheat lines from CIMMYT's Global Wheat Program that had been evaluated under various field-management conditions during more than five cropping seasons and were genotyped with 9045 markers. Some of these lines were also evaluated under the same conditions in South Asia during 2013–2016. A pedigree relationship matrix (***A***) for the 58,798 individuals was computed using the software 'pedigreemm' (Bates and Vazquez, 2014). However, given the dimensions of ***A***, it is difficult to hold it in the computing memory and compute it, so the relationship matrix uses results from partitioned matrices to obtain the result and speed up the computations.

In the study by Pérez-Rodríguez *et al*. (2017), an additional problem was the different sizes of the pedigree matrix ***A*** and the genomic matrix (***G***) because only 29,484 of the wheat lines had been genotyped. The authors showed a single-step model that combined pedigree and marker information into an unified ***H*** matrix by applying the method proposed by Legarra *et al*. (2009) and Aguilar *et al*. (2010), together with a G×E interaction multiplicative model (the reaction norm model of Jarquín *et al*., 2014) with pedigree information (***A***), genomic information (***G***) or both (***H***).

In addition to the drastic increase in the complexity and size of the datasets involved in genomic and pedigree selection, and prediction, massive numbers of genotypes can be screened at a very low cost by using high-throughput phenotyping (HTP) platforms that make use of hyperspectral imaging data. While the main objective of GS is to use massive numbers of markers to reduce phenotyping costs, the aim of HTP is to

have, at a low cost, high-density phenotypes of very large numbers of individuals or breeding lines across time and space using remote or proximal sensing. This can increase both the accuracy and intensity of selection and, therefore, the selection response, while decreasing phenotyping costs. The main idea of HTP is to use predictor traits related to grain yield, disease resistance or end-use quality that may be useful in early-generation testing of lines (Rutkoski *et al.*, 2016; Montesinos-López *et al.*, 2016, 2017a). The CIMMYT Global Wheat Program phenotyped (throughout HTP) and stored hundreds of thousands of wheat field plots in hundreds of thousands of megabytes of data, which are processed and used for selection. Results of using HTP in early-generation testing of wheat lines using canopy temperature, and green and red normalized difference vegetation indexes (GNDVI and RNDVI, respectively) as predictor traits in pedigree and genomic best linear unbiased prediction models could increase prediction accuracy for grain yield (Rutkoski *et al.*, 2016; Montesinos-López *et al.*, 2017a, 2017b).

Therefore, as a result of the advances in GS and genomic prediction, along with the use of HTP at intermediate and advanced stages of the breeding programme, data volumes and complexity have drastically increased, leading to novel research efforts, combining, among other things, computer science, machine learning, mathematics, physics, statistics, genetics, quantitative genetics and bioinformatics. Such work has emerged as a new field of research, commonly known as data science or data-driven science, that aims to unify statistics with data analysis, data mining, machine learning methods and so on. Interdisciplinary researchers in data science focus on computing more accurate predictive values by using statistical models or machine learning models (McDowell, 2016) on big data.

This review covers new theoretical and practical GS advances and outcomes produced in the past 3–4 years in wheat. We first describe new models (improved mixed models and item-based collaborative filtering, or IBCF) that deal with the complexity of genomic-enabled prediction and models for assessing different forms of G×E interaction and marker × environment interaction. We summarize the results of applying GS in real advanced wheat trials, as well as the preliminary results of predicting wheat lines developed in Mexico and evaluated in different environments of South Asia. We also report advances in genomic selection indices and discuss topics related to performing high-throughput phenotyping (phenotyping a large number of individuals) in early-generation testing to accelerate genetic gains.

Materials and Methods

Phenotypic data

The dataset included a total of 45,099 wheat lines that were evaluated at the Norman E. Borlaug Experiment Station in Ciudad Obregon, Mexico, under optimal field-management conditions during five cycles (2013–2018). Table 13.1 contains the number of lines evaluated per year. Original data from each year comprise a large number of trials, each established using an alpha-lattice design with three replicates. The basic model for each year comprises the random effects of trials, the random effects of the replicates within the trials, the random effects of the incomplete blocks within trials and replicates, and the random effects of the breeding lines.

On two dates, two more traits were measured for each genotype: normalized difference vegetation index using green light (GNDVI) and using red light (RNDVI). The two dates were the end of February and beginning of March for cycles 2015–2016, 2016–2017 and 2017–2018. Thus, in total, four traits were measured in each cycle, plus grain yield.

Markers

The genotypic information consisted of 11,293 genotyping-by-sequencing (GBS) markers for 46,089 lines. Lines were genotyped using the

Table 13.1. Number of lines evaluated during five cycles under optimal conditions in Ciudad Obregon, Mexico.

Year	Number of records
2013–2014	7671
2014–2015	9091
2015–2016	9501
2016–2017	9821
2017–2018	9015

Illumina HiSeq2500 sequencer at Kansas State University. Genotyping was performed using the GBS approach (see Poland *et al.*, 2012 and Glaubitz *et al.*, 2014, for more details). Marker polymorphisms were called using TASSEL (https://tassel.bitbucket.io) version 5.0 and the GBS pipeline (Glaubitz *et al.*, 2014) version 2. Figure 13.1 includes the number of SNPs per chromosome within a 1 Mb-size window. Markers with more than 30% missing values were removed. The rest of the markers were imputed using the observed allelic frequencies; after imputing, we computed minor allelic frequencies (MAF) and removed markers with MAF < 0.05. After quality control and imputation, a total of 6978 markers were available for making predictions.

Pedigree

The pedigree for 46,326 individuals was also available. The pedigree was obtained by querying the BROWSE program (McLaren *et al.*, 2000). An additive relationship matrix $\boldsymbol{A}$ for individuals was generated using the pedigreemm package (Bates and Vázquez, 2014); the pedigree takes selfing into account. The development version of the routines for obtaining the relationship matrix, taking the selfing cycles into account, can be obtained from github (https://github.com/Rpedigree/pedigreeTools) and from CRAN. Given the dimensions of matrix $\boldsymbol{A}$, it is difficult to handle it in computing memory. Appendix A in Pérez-Rodríguez *et al.* (2017) shows an R script (R Core Team, 2018) to compute and store the relationship matrix efficiently using partitioned matrixes.

Statistical models

Model 1: genomic G×E interaction using markers

The parametric G×E interaction model takes into account the main effect of E environments, the main effect of genotypes, and the interaction between genotypes and environments. In matrix notation, the model can be written as:

$$\boldsymbol{y} = \mu\mathbf{1} + \mathbf{Z}_E\beta_E + \mathbf{Z}_g\boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{e} \qquad \text{(Eqn 13.1)}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_E)'$ is the response vector, and $\boldsymbol{y}_j$ represents the observations in the j-th environment (j = 1, ..., E). The general mean is μ; $\mathbf{Z}_E$ is an

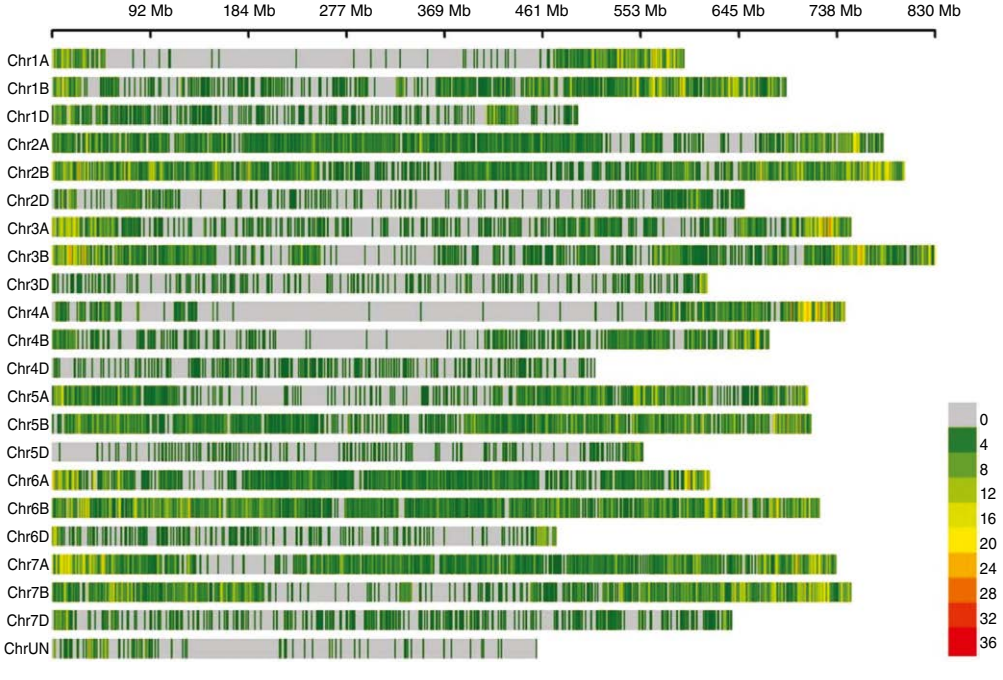


Fig. 13.1. Number of SNPs within a window 1 Mb in size.

incidence matrix for environments, and we assume $\beta_E \sim MN(\mathbf{0}, \sigma_E^2 \boldsymbol{I})$; $\mathbf{Z}_g$ is an incidence matrix that connects genotypes with phenotypes; $\boldsymbol{u}_1$ represents the random effect of genotypes; it is assumed multivariate normal, that is, $\boldsymbol{u}_1 \sim MN(\mathbf{0}, \sigma_u^2 \boldsymbol{G})$; and $\boldsymbol{u}_2$ represents the effect of G×E interaction. We assume $\boldsymbol{u}_2 \sim MN(\mathbf{0}, \sigma_{ge}^2(\boldsymbol{Z}_g \boldsymbol{G} \boldsymbol{Z}_g') \# (\boldsymbol{Z}_E \boldsymbol{Z}_E'))$, where # denotes the Hadamard product (cell-by-cell) of the two matrices in parentheses (see Jarquín *et al.*, 2014; Pérez-Rodríguez *et al.*, 2015). Finally, we assume that the residuals are distributed as follows: $\boldsymbol{e} \sim MN(\mathbf{0}, \sigma_e^2 \boldsymbol{I})$.

Model 2: pedigree G×E interaction

Relationship matrix $\boldsymbol{G}$ is replaced by the relationship matrix derived from pedigree $\boldsymbol{A}$ in model (1). Models were fitted using the bigBGLR package (de los Campos and Pérez-Rodríguez, 2017; https://github.com/gdlc/bigBGLR-R). The bigBGLR package is a fork of the BGLR package (de los Campos and Pérez-Rodríguez, 2017) for dealing with huge matrices in shared memory through the use of the bigmemory package (Kane *et al.*, 2013).

Model 3: item-based collaborative filtering (IBCF)

IBCF is a model-based algorithm for recommender items or products (Montesinos-López *et al.*, 2018a,b). This technique assumes that the data can be arranged in a rectangular format; the rows correspond to 'Users' and the columns correspond to 'Items', and each intersection of a row and column represents the rating of an item given by a specific 'User'. If a user has not rated an item, the rating can be predicted on the basis of the ratings of other 'Users' and other 'Items'. Consider, for example, the case of three users and four items presented in Table 13.2; y_{ij} represents the rating for item j given by user i. The aim here is to predict the rating for 'Item 3' given by 'User 2'.

The problem can be solved by using a weighted average of the rest of the ratings, with the weights computed according to a similarity matrix between 'Items' (Sarwar *et al.*, 2001; Montesinos-López *et al.*, 2018a,b). In the context of genomic prediction in multi-environment trials, the data can also be arranged in a rectangular array, where the rows correspond to genotypes or lines and the columns to specific combinations between traits and environments. Assuming that, for example, three genotypes were evaluated in two environments, and that two traits were evaluated for each genotype, then the data layout would be as shown in Table 13.3.

The data in Table 13.3 are then standardized by columns ($[z_{ij} = (y_{ij} - \mu_j)\sigma_j^{-1}]$), where i denotes the users (lines) and j denotes the columns (trait–environment combinations). With the standardized information, we created a table similar to Table 13.3 with standardized values. The predictions of non-evaluated entries for a given genotype in an environment can be obtained as follows (Montesinos-López *et al.*, 2018b):

$$\hat{y}_{ij} = \hat{\mu}_j + \hat{\sigma}_j \hat{z}_{ij}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ and correspond to the sample mean and sample standard deviation for trait–environment j, calculated with training data, respectively; then

$$\hat{z}_{ij} = \frac{\sum_{j' \in N_i(j)} z_{ij'} w_{jj'}}{\sum_{j' \in N_i(j)} |w_{jj'}|}$$

corresponds to the standardized, predicted phenotype for genotype i in trait–environment j, and $N_i(j)$ denotes the trait–environments for genotype i as being the most similar to trait–environment j.

This works by building a matrix of preferences (called trait-matrix), where each row represents a user (line), and each column represents a trait (item) (four vegetative indices + GY), and the number at the intersection of a row and a column represents the wheat line value for that trait. The absence of a value at this intersection indicates that the line does not have the measured trait. We created a trait-to-trait similarity matrix using

Table 13.2. Rating table.

User/Items	Item 1	Item 2	Item 3	Item 4
User 1	y_{11}	y_{12}	y_{13}	y_{14}
User 2	y_{21}	y_{22}	Not rated	y_{24}
User 3	y_{31}	y_{32}	y_{33}	y_{34}

Table 13.3. Example of a multi-trait, multi-environment data layout.

Genotype/ Trait–Env	T1–E1	T2–E2	T1–E2	T2–E2
Genotype 1	y_{11}	y_{12}	y_{13}	y_{14}
Genotype 2	y_{21}	y_{22}	Not evaluated	y_{24}
Genotype 3	y_{31}	y_{32}	y_{33}	y_{34}

correlations because the idea was to determine how similar one trait is to another. Then, for each line and based on these correlations, we predicted the trait that had not been measured on that line. Predictions were made using the IBCF.MTME package (Montesinos-López *et al*., 2018b). The appendix to this chapter contains the R code for performing predictions using this package.

Assessing prediction accuracy

The main interest here is to predict the performance of non-observed lines for the current/next breeding cycle using historical data. So, the idea here is to predict the grain yield for cycle 2017–2018, using phenotypic and genotypic data from cycles 2013–2014, 2014–2015, 2015–2016 and 2017–2018. Once the predictions are made for the 9015 individuals in the 2018 cycle, the individuals are ranked to select the best yielding lines. The phenotypical response of the 9015 individuals is already known, but the grain yield will be predicted to test the predictive power of the proposed models. The prediction accuracy of the models was assessed on the basis of two criteria: (i) Pearson's correlation between the predicted and observed values, and (ii) a graphical representation of the observed versus the predicted values. Values were sorted into classes based on percentiles of the empirical distribution of observed and predicted values (Jarquín *et al*., 2014).

Results

A boxplot with adjusted grain yield, by breeding cycle, for lines evaluated under optimal conditions is shown in Fig. 13.2. One of the goals of a breeding programme is to have lines with high grain yield, but Fig. 13.2 reveals that there is natural variation in grain yield from cycle to cycle because of the effect of environmental conditions and also because the evaluated materials are changed from cycle to cycle. Table 13.4 contains the Pearson correlation coefficients between observed and predicted values for models

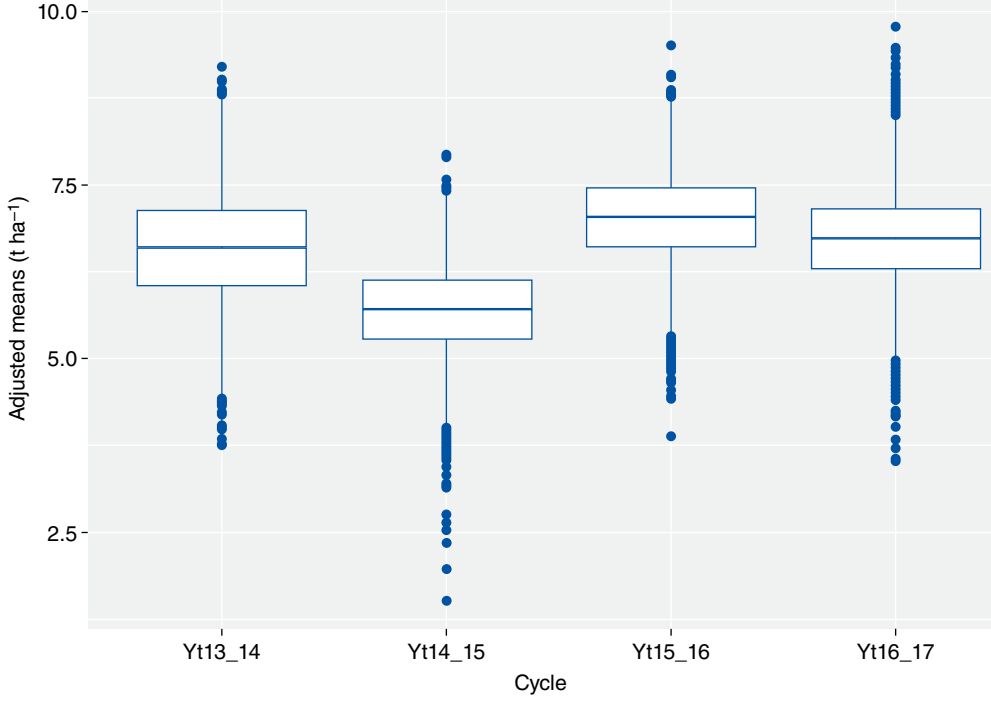


Fig. 13.2. Boxplot of adjusted grain yield by cycles.

based on markers, pedigree, markers + pedigree and IBCF. Note that the highest correlation is for a model based on markers + pedigree, that is, the simple average between the predictions obtained with the model based on markers and the model based on pedigree. Note also that the predictions based on IBCF are very close to the correlations based on pedigree only. Figure 13.3 represents a scatter plot of predictions based on markers + pedigree against the adjusted means. Figure 13.4 represents a scatter plot of predictions based on IBCF versus adjusted means. It is clear from these plots that the mean squared error of predictions based on markers and pedigree is lower than that based on IBCF, but predictions based on IBCF are based only on a few covariates.

We took the top 2000 individuals of the observed lines (adjusted phenotypic means in Yt2017–2018) in the training set (TRT) and the top 2000 predicted lines of the best predictive model – ½(marker + pedigree model) – in the testing set (TST). A total of 948 lines of the top 2000 lines in the TST belong to the top 2000 lines in the observed values of the TRT. Thus, the match percentage = (948/2000) × 100 = 47.4% (Fig. 13.5).

Discussion

In the era of Big Data, data acquisition in all areas of science is growing faster than the ability to store, distribute and analyse data to extract

Table 13.4. Correlations in the 2018 testing set (9015 wheat lines).

Comparison	Pearson's correlation
Markers versus adjusted means GY testing 2017–2018	0.4263
Pedigree versus adjusted means GY testing 2017–2018	0.3547
1/2(markers + pedigree) versus adjusted means GY testing 2017–2018	0.4374
IBCF_2018_a versus adjusted means GY testing 17–18 (phenotypic data from 2015–2016 and 2016–2017 in the training set to predict 2017–2018 data)	0.3763
IBCF_2018_b versus adjusted means GY 2017–2018 (phenotypic data from 2016–2017 in the training set to predict Yt2017–2018 data)	0.3614

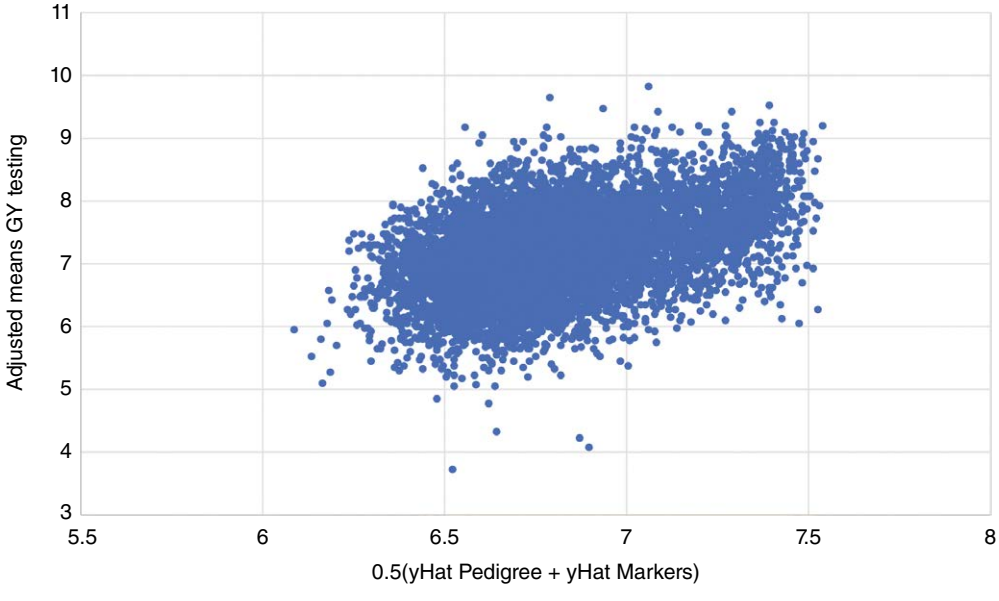


Fig. 13.3. Predictive values based on ½ (pedigree + marker) versus adjusted means GY testing, (2017–2018); Pearson's correlation = 0.4374.

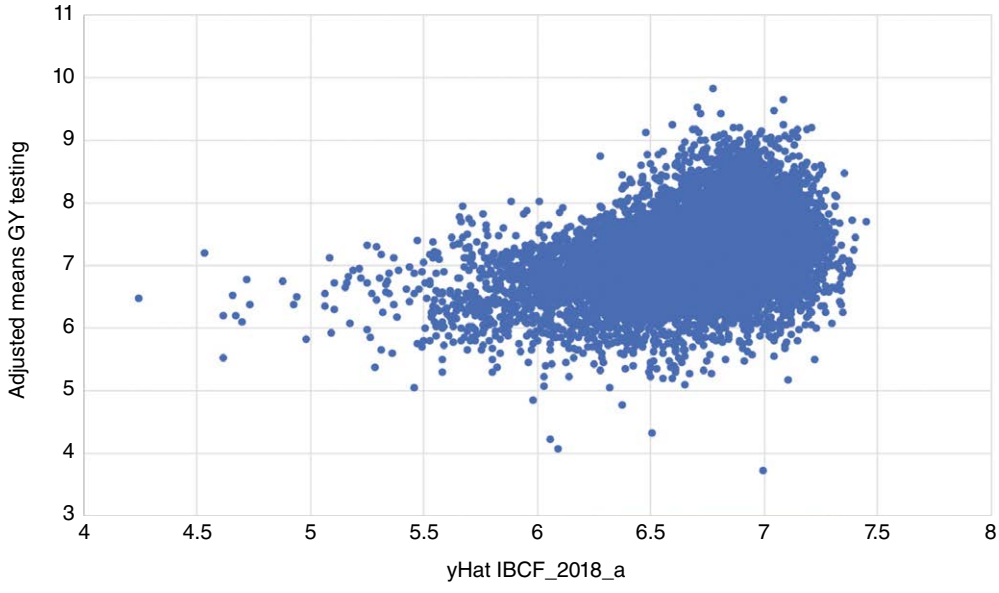


Fig. 13.4. Predictive values based on IBCF_2018_a versus adjusted means GY testing 2017–2018; Pearson's correlation = 0.3763.

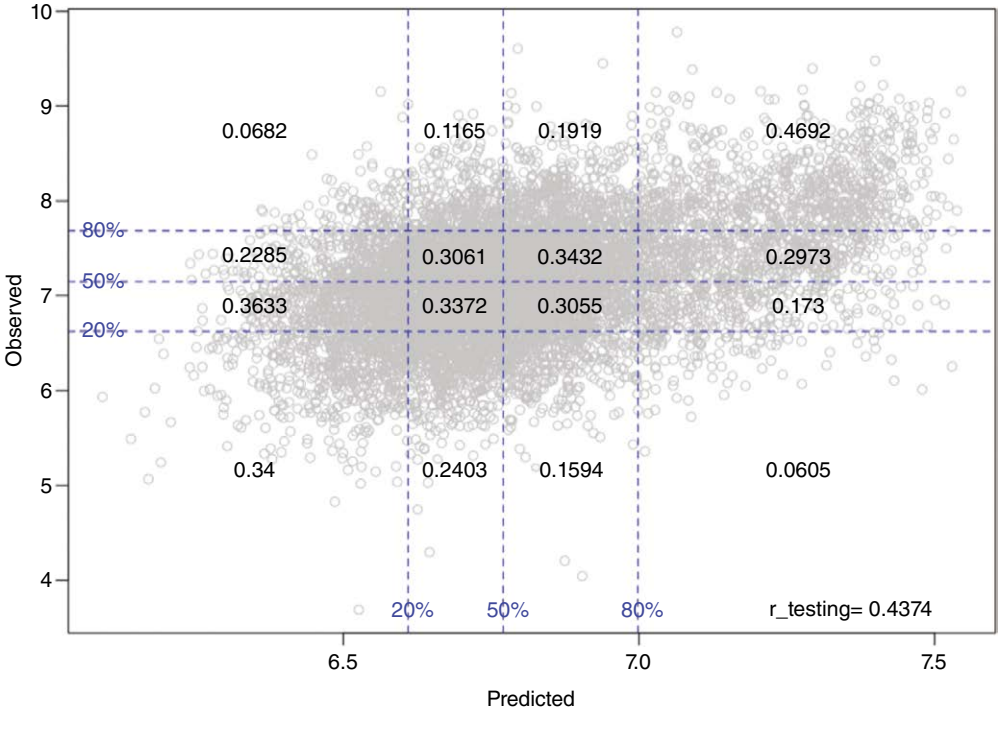


Fig. 13.5. Conditional plot of predicted values (average of predictions of model based on markers and pedigree) versus observed values (adjusted means GY testing Yt2017–2018).

useful knowledge/information. For example, the sequencing of the human genome – i.e. determining the complete sequence of the 3 billion DNA base pairs and identifying each human gene – took 13 years (from 1990 to 2003) and required an investment of US$3.8 billion; however, nowadays sequencing a human genome takes 1 to 2 days and costs less than US$1500 (Tripp and Grueber, 2011). Moreover, Stephens *et al*. (2015) predicted that by 2025, between 100 million and 2 billion human genomes will have been sequenced, with data storage demands of 2 to 40 exabytes. That's more than the projected storage needs of YouTube and Twitter. Stephens *et al*. (2015) also stated that genomics is a Big Data science that will continue to get much bigger; for this reason, it is considered the 'four-headed beast' (after Twitter, YouTube and astronomy), since it will pose some of the most severe computational challenges in the coming years or decades relative to the life cycle of a dataset – acquisition, storage, distribution and analysis.

Among the breeding methods used to accelerate the release of new genotypes, GS is currently the most promising one. Several statistical models, based on the standard GBLUP that incorporates G×E interactions in genomic and pedigree predictions, have shown substantial increases in the prediction accuracy of individuals unobserved in test environments. These GS prediction models can help scientists in different disciplines develop drought- and heat-tolerant plants by exploiting positive G×E interactions. For GS, a plethora of new statistical methods have been developed for predicting unobserved individuals. In general, machine learning algorithms and methods have been very successful in recognizing complex patterns and making correct decisions based on data. Kernel-based methods, such as reproducing Hilbert spaces regression, have extensively delivered good genomic predictions in plants.

New integrative approaches need to be developed that take into account challenges in all four aspects; it is unlikely that a single advance or technology will solve the genomics data problem. For this reason, there is an active area of research that seeks to improve the existing statistical models to deal better with large datasets and increase their prediction accuracy. For this reason, in this chapter, we compared two approaches (improved mixed models and IBCF) for dealing with moderately large datasets in the context of GS. We found that under the improved mixed statistical models, the prediction accuracy based on markers (0.4263) was higher than the prediction accuracy based on pedigree (0.3547). We also found that prediction accuracies were, in general, higher than those obtained in previous cycles. On the other hand, we observed that the IBCF method made good predictions based only on phenotypic data and four covariates of HTP data. This quality of the IBCF based on correlated traits should be better exploited. One of the great advantages of the IBCF is that it is fast and provides very high prediction accuracies when used with a large number of correlated variables. However, it is also important to point out that when the target trait (to be predicted) and the covariates assumed to be known in the IBCF are weakly correlated, the performance of the IBCF is poor. For this reason, it is of paramount importance to be very careful when implementing the IBCF. Finally, the improved mixed statistical models generally showed better performance, which may be attributable to the fact that they take into account marker, pedigree information and G×E interactions, which allows these models to borrow information from correlated lines across correlated years. For this reason, the combination of information on pedigree, genomics and G×E interaction gave the best prediction accuracies.

Finally, as previously stated, we are aware that to meet the four big challenges of genomics, we need to work more collaboratively. More people need to work on these challenging aspects. Although many people in other areas of science are working on similar things, the solutions needed here are domain specific. It is also necessary to prioritize aspects that are growing more slowly, such as analyses that turn data into knowledge in a more reliable and efficient way.

Summary and Concluding Remarks

In this chapter, we compared two approaches for genome-based prediction. First, we provided a general overview of the challenges we face in the context of GS in the era of big data. Then we described and implemented with real datasets two methods of dealing with the problem of using moderate datasets in GS to select the best

individual candidates early. We compared these two methods relative to prediction accuracy and pointed out the advantages and disadvantages of each. Finally, we addressed the need to continue making collaborative efforts and conduct more scientific research to improve, in the GS context, the lifecycle of a dataset by putting more emphasis on data analyses that are essential for turning data into useful knowledge.

G×E interaction plays an important role in the selection of plant materials with important agronomic characteristics (e.g. high grain yield and resistance to several diseases), which are well adapted to different growing conditions. Traditionally, plant breeding has been performed using phenotypic records obtained from field evaluations and with the help of pedigree records. Almost two decades ago, with the introduction of genomic selection based on dense molecular markers, the breeding process underwent a revolution. Modern breeding programmes are also able to register high dimensional environmental covariates and, in some cases, it is also possible to incorporate information from hyperspectral imaging cameras through the generation of vegetation indexes or the use of high dimensional wavelength data. However, incorporating all these sources of high dimensional data is not easy, because it requires the use of appropriate statistical tools and efficient computer algorithms implemented in modern computing languages. In this work, we analyse high dimensional data from CIMMYT's wheat breeding programme, which includes more than 45,000 wheat lines that were genotyped using dense SNP markers and have existing pedigree records. Some hyperspectral images are available for some of the breeding cycles. We predicted the performance of unobserved lines using linear models that incorporate markers, pedigree, and the interaction between genotype and environment, as well as a new approach that makes predictions based on recommender systems that are routinely used in e-commerce, marketing, biology and, fairly recently, in genomic selection.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., *et al*. (2010) Hot topic: A unified approach to utilise phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93, 743–752.

Bates, D. and Vazquez, A.I. (2014) pedigreemm: Pedigree-based mixed-effects models. R package version 0.3-3. Available at: https://CRAN.R-project.org/package=pedigreemm (accessed 7 May 2019).

Bernardo, R. and Yu, J.M. (2007) Prospects for genome-wide selection for quantitative traits in maize. *Crop Science* 47, 1082–1090.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science* 52, 707–719.

Crossa, J., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Burgueño, J., *et al*. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724.

Crossa, J., Pérez-Rodríguez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., *et al*. (2011) Genomic selection and prediction in plant breeding. *Journal of Crop Improvement* 25, 239–226.

Crossa, J., Beyene, Y., Kassa, S., Pérez-Rodríguez, P., Hickey, J.M., *et al*. (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes, Genomes*, *Genetics* 3, 1903–1926.

Crossa, J., Pérez-Rodríguez, P., Hickey, J., Burgueño, J., Ornella, L., *et al*. (2014) Genomic prediction in CIMMYT maize and wheat breeding programmes. *Heredity* 112, 48–60.

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., *et al*. (2016) Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics* 6(7), 1819–1834. DOI: 10.1534/g3.116.029637.

Daetwyler, H.D., Kemper, K.E., van der Werf, J.H.J. and Hayes, B.J. (2015) Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of Animal Science* 90, 3375–3384. DOI: 10.2527/jas2011-4557.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., *et al*. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385.

de los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K. and Crossa, J. (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* 92, 295–308.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. and Calus, M.P.L. (2012) Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2), 327–345. DOI: 10.1534/genetics.112.143313.

de los Campos, G., Veturi, Y., Vazquez, A.I., Lehermeier, C. and Pérez-Rodríguez, P. (2015) Incorporating genetic heterogeneity in whole-genome regressions using interactions. *Journal of Agricultural, Biological, and Environmental Statistics* 20, 467–490. DOI: 10.1007/s13253-015-0222-5.

de los Campos, G. and Pérez-Rodríguez, P. (2017) bigBGLR: Bayesian Generalized Linear Regression. R package version 1.0.5. Available at: https://github.com/gdlc/bigBGLR-R (accessed 1 April 2019).

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., *et al*. (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9, e90346. DOI: 10.1371/journal.pone.0090346.

González-Camacho, J.M., de los Campos, G., Pérez-Rodríguez, P., Gianola, D., Cairns, J., *et al*. (2012) Genome-enabled prediction of genetic values using Radial Basis Function Neural Networks. *Theoretical and Applied Genetics* 125(4), 759–771. DOI: 10.1007/s00122-012-1868-8.

González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, O. and Gianola, D. (2016) Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics* 17, 208. DOI: 10.1186/s12864-016-2553-1.

Heslot, N., Yang, H.P., Sorrells, M.E. and Jannink, J.L. (2012) Genomic selection in plant breeding: A comparison of models. *Crop Science* 52, 146–160.

Heslot, N., Akdemir, D., Sorrells, M.E. and Jannink, J.L. (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics* 127, 463–480.

Hickey, J.M., Crossa, J., Babu, R. and de los Campos, G. (2012) Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science* 52, 654–663.

Janss, L., de los Campos, G., Sheehan, N. and Sorensen, D. (2012) Inferences from genomic models in stratified populations. *Genetics* 192(2), 693–704.

Jarquín, D., Crossa, J., Lacaze, X., Cheyron, P.D., Daucourt J., *et al*., (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127, 595–607.

Kane, M.J., Emerson, J.W. and Haverty, P. (2013) bigmemory: Manage massive matrices with shared memory and memory-mapped files. R package version 4.4.5. Available at: http://CRAN.R-project.org/package=bigmemory (accessed 1 April 2019).

Legarra, A., Aguilar, I. and Misztal, I. (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* 92(9), 4656–4663.

López-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., *et al*. (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *G3: Genes, Genomes, Genetics* 5(4), 569–582. DOI: 10.1534/g3.114.016097.

Lorenzana, R.E. and Bernardo, R. (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120, 151–161.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

McDowell, R.M. (2016) Genomic selection with deep neural networks. MSc dissertation, Iowa State University, Ames, Iowa, USA. DOI: 10.31274/etd-180810-5600.

McLaren, C.G., Ramos, L., López, C. and Eusebio, W. (2000) Applications of the genealogy management system. In: McLaren, C.G., White, J.W. and Fox, P.N. (eds) *International Crop Information System. Technical Development Manual*, version VI. CIMMYT and IRRI, Mexico City, Mexico, pp. 5.8–5.13.

Montesinos-López, A., Montesinos-López, O.A., Crossa, J., Burgueño, J., Eskridge, K., *et al*. (2016) Genomic Bayesian prediction model for count data with genotype × environment interaction. *G3: Genes, Genomes, Genetics* 6(5), 1165–1167. DOI: 10.1534/g3.116.028118.

Montesinos-López, A., Montesinos-López, O.A., Cuevas, J., Mata-López, W.A., Burgueño, J., *et al*. (2017b) Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. *Plant Methods* 13(62), 1–29.

Montesinos-López, O.A., Montesinos-López, A., Pérez-Rodríguez, P., de los Campos, G., Eskridge, K.M., *et al*. (2015a) Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. *G3: Genes, Genomes, Genetics* 5(1), 291–300.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Burgueño, J. and Eskridge, K. (2015b) Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. *G3: Genes, Genomes, Genetics* 5(10), 2113–2126. DOI: 10.1534/g3.115.021154.

Montesinos-López, O.A., Montesinos-López, A., Pérez-Rodríguez, P., Eskridge, K., He, X., *et al*. (2015c) Genomic prediction models for count data. *Journal of Agricultural, Biological, and Environmental Statistics* 20(4), 533–554.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., de los Campos, G., Alvarado, G., *et al*. (2017a) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* 13(4), 1–23.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Montesinos-López, J.C., Mota-Sánchez, D., *et al*. (2018a) Prediction of multiple-trait and multiple-environment genomic data using recommender systems. *G3: Genes, Genomes, Genetics* 8, 131–147.

Montesinos-López, O.A., Luna-Vázquez, F.J., Montesinos-López, A., Juliana, P., Singh, R., *et al*. (2018b) An R package for multitrait and multienvironment data with the item-based collaborative filtering algorithm. *The Plant Genome* 11(3), 1–16.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manes, Y., *et al*. (2012) Comparison between linear and non-parametric models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genetics* 2, 1595–1605.

Pérez-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., *et al*. (2015) A pedigree reaction norm model for prediction of cotton (*Gossypium* sp.) yield in multi-environment trials. *Crop Science* 55, 1143–1151.

Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., *et al*. (2017) Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *The Plant Genome* 10(2) 1–15. DOI: 10.3835/plantgenome2016.09.0089.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S.Y., *et al*. (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. DOI: 10.3835/Plantgenome2012.06.0006.

R Core Team. (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at: https://www.R-project.org/ (accessed 1 May 2019).

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., *et al*. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* 44(2), 217–220.

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., González-Pérez, L., *et al*. (2016) Canopy temperature and vegetation indices from high throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics* 6(9), 2799–2808.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001) Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the 10th International Conference on the World Wide Web*, ACM, New York, pp. 285–295.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., *et al*. (2015) Big data: Astronomical or genomical? *PLoS Biology* 13(7), e1002195. DOI: 10.1371/journal.pbio.1002195.

Technow, F., Bürger, A. and Melchinger, A.E. (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes, Genomes, Genetics* 3(2), 197–203.

Tripp, S. and Grueber, M. (2011) *Economic Impact of the Human Genome Project*. Available at: https://www.battelle.org/docs/default-source/misc/battelle-2011-misc-economic-impact-human-genome-project.pdf (accessed 1 March 2019).

Windhausen, V.S., Atlin, G.N., Crossa, J., Hickey, J.M., Grudloyma, P., *et al*. (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3: Genes, Genomes, Genetics* 2(11), 1427–1436. DOI: 10.1534/g3.112.003699.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H.P., *et al*. (2012) Accuracy of genomic selection in European maize elite breeding populations. *Theoretical and Applied Genetics* 124(4), 769–76.

Appendix

This appendix contains the R code for obtaining predictions based on IBCF. The functions are included in the R library IBCF.MTME. The data are read from a csv file. The predictions are made using the R function 'IBCF.Years', which uses information from other years as training. The function takes the following arguments:

- Dataset: a data frame with a column for years, a column that identifies the genotypes and the rest of the columns show the traits evaluated for each genotype in each year.
- colYears: the id or position of the column that specifies the year.
- Years.testing: the id of the year to be predicted.
- Traits.testing: the id of the trait to be predicted.

Once the prediction task is performed, the function returns an object with several outputs, the most important being a vector denominated 'predicted', with the predicted values for the target trait. The 'IBCF.Years' function was extensively documented in Montesinos-López *et al.* (2018b). Below we show the R code used for obtaining the predictions for cycle 2017–2018.

```
#Clean workspace
rm(list = ls())

#Load library IBCF.MTME
library(IBCF.MTME)

################Loading the data#############################
Dataset=read.csv(file="YTBWY_ALL_NDVI_GY_M.csv",
                 header=TRUE,na.strings="NA")

Dataset$GID=paste(Dataset$TrialNo,Dataset$GID,sep="_")

#Remove TrialNo column
Dataset=Dataset[,-2]

# Remove unnamed Year
#(or could add a name and don't remove it)
Dataset=Dataset[which(!is.na(Dataset$Year)), ]

#Individuals to be predicted in 2018
index_2018=which(Dataset$Year==2018)

#Obtain predictions based on IBCF

out_IBCF=IBCF.Years(Dataset, colYears = 'Year',
                    Years.testing = c(2018),
                    Traits.testing = c('GY'))

#Obtain the predictions for all the Years
predictions_all=out_IBCF$predicted

#Obtain the predictions for Year 2018 (Cycle 2017-2018)
predictions_2018=predictions_all[index_2018]
```

```
#Add a column to the input Dataset
Dataset$GY_Pred_2108=predictions_all

#Write the results
write.csv(Dataset,
          file="Precited_YTBWY_ALL_NDVI_GY_2018_Final.csv")
```

14 Quantitative Genetics in Improving Root and Tuber Crops

Hernán Ceballos*
International Center for Tropical Agriculture (CIAT), Cali, Colombia

Introduction

Several genetic designs have been developed and used extensively during the second half of the 1900s. The reproductive biology of crops influences the type of design preferred by researchers. In cross-pollinated crops, such as maize, diallels and North Carolina Designs are often used, whereas generation mean analysis is generally preferred by breeders working with self-pollinated crops. The development of such designs has been highly influential in improving grain crops. Quantitative genetic research has also been conducted in root and tuber crops, which are vegetatively propagated, but not as extensively as in grain crops. The special needs and opportunities of asexually reproduced crops are currently addressed by the Research Program on Roots, Tubers and Bananas (RTB) of the Consultative Group on International Agricultural Research (CGIAR).

The RTBs offer special advantages and problems in relation to the use of genetic designs for the analysis of quantitatively inherited traits. For example, their clonal multiplication allows evaluating the same genotype across repetitions and/or locations. The genetic variation among siblings within a given family can thus be separated from the micro-environmental conditions associated with the growth of each individual plant. This is not feasible in sexually propagated crops, unless progenitors are inbred. In that case, however, there is no within-family genetic variation. Measuring the genetic variation within families, in turn, allows a test of epistasis via analysis of diallel crosses (Hallauer and Miranda, 1988, p. 59), which is not possible in grain crops.

Typically, the subjects targeted by quantitative genetic analyses in grain crops are the families derived from a random or fixed set of progenitors. Little or no attention is paid to individual genotypes within a family and their performances are pooled together into family output. In the case of RTBs, on the other hand, breeders pay special attention to individual genotypes in search of those with outstanding performance. Superior genotypes are then clonally multiplied and, eventually, released as varieties. These major differences distinguishing grain crops from RTB crops form the basis of the present chapter.

Self- or cross-incompatibility is of common occurrence in certain RTBs and imposes a limitation in the implementation of certain genetic designs. The RTBs are often polyploid (Table 14.1), which complicates breeding and quantitative genetic analyses considerably. Elimination of deleterious alleles in polyploid species, for example, is quite difficult (Jansky and Spooner, 2018). Inbreeding depression is a common feature in different RTBs,

* Email: H.CEBALLOS@cgiar.org

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

Table 14.1. Relevant characteristics of different root and tuber crops.

Crop	Ploidy	Sexual reproductive biology
Cassava (*Manihot esculenta* Crantz)	2x	No compatibility problem. Only three seeds per cross. Functional diploid (Nassar *et al*., 2010; Wang *et al*., 2011).
Potato (*Solanum tuberosum* L.)	4x	Compatibility problems. Many seeds per cross (Bradshaw, 2007).
Wild potatoes	2x–6x	
Sweet potato (*Ipomoea batatas* L.)	6x	Self-pollination difficult. Four seeds per cross (Tan *et al*., 2007; Lebot, 2010a).
Greater yam (*Dioscorea alata*)	4x	Dioecious reproduction leads to either male or female genotypes. Major problem in obtaining seeds as flowering is often scarce (Gamiette *et al*., 1999; Dansi *et al*., 2001; Bousalem *et al*., 2006).
White guinea yam (*Dioscorea rotundata*)	4x–8x	
Yellow guinea yam (*Dioscorea cayenensis*)	4x–8x	
Aja, cush-cush, yampi (*Dioscorea trifida*)	4x	
Taro (*Colocasia esculenta* L.)	2x–3x	Restricted as often genotypes do not flower. Gibberellins promote flowering in some cases. Lack of synchrony (McDavid and Alamu, 1976; Volin and Beale, 1981; Wilson, 1990; Iramu *et al*., 2010; Oumar *et al*., 2011; Amadi *et al*., 2015; Obidiegwu *et al*., 2016).
Cocoyam (*Xanthosoma sagittifolium* L.)	2x–4x	

which in turn implies that heterosis and non-additive genetic effects are often important as well (Mendoza and Haynes, 1974; Easwari Amma *et al*., 1995; Bradshaw, 2007; Lin *et al*., 2007; Chaurasiya *et al*., 2013; Chipeta *et al*., 2013; Gopal, 2014; Ceballos *et al*., 2015; Mwanga *et al*., 2017; Gurmu *et al*., 2018). Evidence and/or report of heterosis in yams, taro and cocoyam, however, is limited (Ivancic and Lebot, 2000; Quero-García *et al*., 2009).

Sexual reproduction in the RTBs is not as straightforward as it is in the case of grain crops, which depend heavily on it for their survival. Often, genotypes from different RTBs fail to flower completely or do so scarcely (Table 14.1). Synchronization of flowering among different genotypes is often a problem. Obtaining a balanced set of botanical seeds from the crosses that most genetic designs require, therefore, may be challenging in the case of the RTBs.

The sexual reproduction in the RTBs has evolved to typically favour cross pollination; thus, from the genetic point of view, RTB varieties are hybrids that are clonally multiplied. Progenitors and their resulting progenies are usually highly heterozygous, which allows for the preservation of a sizable genetic load (Grüneberg *et al*., 2015; Ramu *et al*., 2017; Jansky and Spooner, 2018; Manrique-Carpintero *et al*., 2018). The asexual reproduction in the RTBs also results in a lower rate of genetic recombination across time compared with grain crops (Ortiz, 1997; de Meeûs *et al*., 2007). Higher rates of asexual reproduction increase heterozygosity and decrease population differentiation. Diversity at single loci is higher in clonal crops compared with those with strict sexual reproduction. The opposite is true, however, for genotypic diversity (Balloux *et al*. 2003).

Another common limitation in the RTBs is the tendency to have poorly developed population structure (or reports about it) compared with grain crops. No heterotic patterns have been identified in the RTBs. While genetic diversity is essential for heterosis, genetic distances have been proven to be inefficient for the identification of heterotic groups (Chołuj *et al*., 2014; Gopal, 2014; Ceballos *et al*., 2016a). There are several reasons for the poor heterosis-predictive capacity of genetic distances and they are relevant to understand the foundations of this

important phenomenon. Genetic distances cannot differentiate different regions of the genome and, in this context, it is assumed that loci involved in heterosis are uniformly distributed. However, this is not the case, as demonstrated decades ago (Stuber *et al.*, 1992; Stuber, 1994; Frascaroli *et al.*, 2007). Transcriptome-based distances have shown better correlations with heterosis than conventional genetic distances (Frisch *et al.*, 2010), suggesting that gene expression and epigenetic factors influence heterosis. The confounding effect of epistasis also contributes to the limited capacity of genetic distance to explain heterosis. Epistasis has been demonstrated to be important for fresh root yield in cassava (Cach *et al.*, 2005; Pérez *et al.*, 2005a,b). An important feature of the RTBs is that their vegetative multiplication allows maintaining epistatic relationships among loci, thus they can be exploited by farmers growing clones that have been found to be outstanding. However, positive epistatic interactions are broken when the same superior clones are used by breeders in crossing blocks.

A distinctive feature of the RTBs is that they are among the only major staple crops whose breeding is based on the use of heterozygous progenitors (except for the recent, revolutionary and successful efforts made in potato), which complicates considerably the development of useful genetic maps based on molecular markers. Heterozygous progenitors also result in within-family genetic variation, a feature that is unusual in grain crops, which rely, more often than not, on inbred progenitors. The relative magnitude of within-family genetic variation, combined with the attention that RTB breeders pay to individual genotypes (rather than to the families they belong to), results in key parameters, such as breeding values of progenitors that play a minor role in RTB breeding compared with sexually propagated crops (Ceballos *et al.*, 2016b).

Breeding Methods, Inbreeding Depression, Epistasis and Heterosis in RTBs

All RTB crops are currently bred through phenotypic mass selection and take advantage of their vegetative propagation (Grüneberg *et al.*, 2015; Ceballos *et al.*, 2017a; Jansky and Spooner, 2018). However, some modifications have been introduced or attempted across years. Developing a system to exploit heterosis through a sort of reciprocal recurrent selection has been implemented in sweet potato (Grüneberg *et al.*, 2015); the use of breeding value of cassava progenitors was suggested by Ceballos and co-workers in 2004; and a shift toward the use of inbred progenitors in potato was recently reported by Jansky and Spooner (2018). In selecting outstanding clones, all genetic effects (additive, dominance and epistatic) are exploited by farmers (Jennings and Iglesias, 2002). However, most of the conventional recurrent selection systems lack the capacity to direct genetic improvement in such a way that the frequency of favourable genetic combinations (within or between loci) is maximized. In other words, while the genetic superiority of outstanding clones can be exploited fully by farmers, a considerable proportion (e.g. non-additive genetic effect) is lost as soon as they are used as progenitors in breeders' nurseries.

Heterosis

Heterosis is the superiority of the hybrid over the average of its two inbred progenitors (Falconer, 1981). However, the difference between the performance of a cross and the average of its non-inbred progenitors is also considered heterosis. This wider definition of heterosis is the one most relevant for the RTBs. Heterosis of a cross between two inbred lines is generally much larger than when non-inbred progenitors are used as parents. The most outstanding and earlier examples of the exploitation of heterosis represent the work done in different grain crops, such as maize, sorghum and sunflower (Stuber, 1994; Duvick, 1999; Troyer, 2006). Commercial hybrids of many vegetable crops (rapeseed, tomato, onion, aubergine, pepper, etc.), sugar beet and cotton followed suit. Surprisingly, heterosis is also exploited in self-pollinated crops, such as rice and even wheat that show interesting levels of heterosis (Fu *et al.*, 2014). Heterosis is therefore a common phenomenon in many different types of crops, belonging to different taxa, and following diverse reproductive systems.

As already mentioned, several reports have highlighted the importance of heterosis in different RTB crops (Mendoza and Haynes, 1974;

Easwari Amma *et al.*, 1995; Bradshaw, 2007; Lin *et al.*, 2007; Chaurasiya *et al.*, 2013; Chipeta *et al.*, 2013; Gopal, 2014; Ceballos *et al.*, 2015; Grüneberg *et al.*, 2015; Mwanga *et al.*, 2017; Gurmu *et al.*, 2018). In fact, many of these articles refer to the importance of non-additive genetic effects, rather than heterosis itself. However, in contrast to crops lacking vegetative reproduction, very little progress has been made (or reported) in the development of breeding methods to exploit heterosis in RTB crops.

Despite the great importance of heterosis, its genetic foundations are still unclear. Two prominent theories (dominance and overdominance) were developed a century ago to explain heterosis. After years of controversy and debate, it is now generally accepted that additive and dominance effects are the main components of heterosis. The possible roles of overdominance and epistasis are still uncertain (Lamkey and Edwards, 1999; Crow, 2000). Molecular technologies have been used to demonstrate the importance of epistasis in the expression of heterosis (Melchinger *et al.*, 2007). Goff postulated in 2011 a cell-based quality control mechanism that detects and downregulates alleles encoding unstable proteins in hybrids. According to this author, it is now feasible to identify alleles encoding unstable proteins and use molecular breeding to eliminate highly expressed alleles encoding unstable proteins. A modification of Goff's model has been suggested (Zhang *et al.*, 2016) after demonstrating that heterosis of complex quantitative traits is influenced by the heterosis of its inherent component traits.

Heterosis is clearly related to the level of heterozygosity. In polyploid species, the frequency of heterozygosity is much higher than in diploid species. For example, the frequency of heterozygosity for allelic frequencies of $p = q = 0.5$ in a diploid species is also 0.50, but above 0.80 in tetraploid and above 0.95 in hexaploid species (Grüneberg *et al.*, 2015). Higher frequency of heterozygosity is also favoured by the vegetative reproduction typical of RTB crops (Balloux *et al.* 2003; de Meeûs *et al.* 2007). The expected levels of heterozygosity contrast drastically with recent results that suggest high levels of homozygosity in cassava (Ramu *et al.*, 2017; Wolfe *et al.*, 2019). Heterotic responses in polyploid species, therefore, show a different pattern compared with those observed in diploid ones. Rather than reaching a maximum in single-cross hybrids, heterosis increases with subsequent generation of crosses. Groose and co-workers (1989) referred to it as *progressive heterosis*. Similar observations have been made in potato (Sanford and Hanneman, 1982; Bani-Aameur *et al.*, 1991) and polyploids in general (Goff, 2011).

Inbreeding

Inbreeding depression is the converse of heterosis, opposite sides of the same coin, as they are often described. Charles Darwin reported as early as 1876 inbreeding depression in different crops, including maize. Several authors have emphasized the importance of understanding inbreeding depression, regardless of the levels of ploidy, in designing efficient breeding methods (Jones and Bingham, 1995; Lamkey and Edwards, 1999). Wright proposed in 1922 that, in the absence of significant linkage or epistasis, there should be a linear relationship between inbreeding depression and inbreeding coefficient (f). Several studies have confirmed this relationship across years, mostly in maize (Robinson and Cockerham, 1961; Hallauer and Sears, 1973; Cornelius and Dudley, 1974; Burton *et al.*, 1978; Sprague, 1983; Lamkey and Edwards, 1999). Lamkey and Edwards summarized in 1999 the main features of inbreeding depression (for diploid species) as follows:

1. A locus will not contribute to inbreeding depression if dominance effects are negligible.
2. The direction of change in the average is toward the value of the recessive allele.
3. Inbreeding depression is maximized when $p = q = 0.5$ (where p and q are the frequencies of the two possible alleles in a diploid species).
4. In the absence of epistasis, inbreeding is a linear function of f.
5. If there is epistasis but no dominance, there will not be any inbreeding depression.
6. If there is epistasis and dominance, then inbreeding depression will be a quadratic or higher function of f.

The most efficient approach to achieving homozygosity is through successive self-pollinations or through the production of doubled haploids. However, in many RTBs, self-incompatibility often occurs. When that is the

case, inbreeding can be attained by crossing related genotypes but at a much slower rate.

There is no problem self-pollinating cassava. In fact, the genotype used for the first sequencing of the cassava genome was an S_3 partial inbred (AM 560-2). Early work to develop homozygous cassava was interrupted because in the process early-branching (e.g. early-flowering) genotypes were favoured and this resulted in an unacceptable plant phenotype (Ceballos *et al.*, 2015). In theory, however, there is no biological barrier to producing fully inbred cassava. Recent advances inducing flowering through grafting, extended photoperiod, use of plant growth regulators and/or pruning young branches can overcome the initial problems in producing homozygous cassava through successive self-pollinations (Ceballos *et al.*, 2017b; Pineda *et al.*, 2018a–d). Dewey stated in 1966 (p. 144) that 'the full potential and effectiveness of a plant-breeding program can be realized only when the reproductive, genetic, and breeding behavior of a species is known and understood'. Current developments in cassava are precisely contributing not only to our understanding but also towards manipulating the reproductive biology of cassava. Therefore, more efficient breeding and deeper knowledge of the genetics of the crop can be attained.

Perhaps, the most convincing evidence of the occurrence of heterosis in the RTBs arises from the strong inbreeding depression in these crops. Several articles have measured inbreeding depression in cassava (Contreras Rojas *et al.*, 2009; Kawuki *et al.*, 2011; de Freitas *et al.*, 2016, 2017). In general, inbreeding depression is high for traits such as fresh root yield (>60%), but not so high for traits such as dry matter content (5%) and plant height (10%). Inbreeding depression, therefore, tends to be higher in traits where non-additive genetic effects are predominant (Ceballos *et al.*, 2015).

Breakthrough research is currently ongoing on the use of inbred progenitors in potato, which relies on the use of diploid germplasm. As early as 1971, de Jong and Rowe had used self-fertile crosses between diploid species from Phureja and Stenotomum groups and haploids from the Tuberosum group. Today, true seed hybrids from inbred progenitors are being tested for commercial production (Lindhout *et al.*, 2018). Figure 14.1 illustrates a large field of inbred potato progenitors used to that end. Inbreeding potato resulted, as expected, in strong inbreeding depression, particularly during the initial cycles of the process. However, as done with temperate maize a century ago or tropical maize in the 1970s, a few cycles of recurrent selection helped build up tolerance to inbreeding (Contreras Rojas *et al.*, 2009).

Most sweet potato genotypes will not produce seed by self-pollination (Martin, 1987). However, it has been reported that, in some genotypes, it was possible to obtain self-pollinated seed, which led to clear inbreeding depression in the resulting seedling plants (van Rheenen, 1964; Komaki *et al.*, 1998). Inbreeding depression in sweet potato was around 50% for yield and 6% for dry matter content. These inbreeding

Fig. 14.1. The implementation of a revolutionary approach for potato breeding through the use of inbred progenitors. (a) Greenhouse where crosses and self-pollinations are made. (b) A large field planted with inbred potato. (Courtesy of Solynta, Inc.)

depression values were the averages across S_1 families from 11 progenitors (Komaki *et al.*, 1998). Interestingly, inbreeding depression levels for yield and dry matter content in sweet potato are similar to those in cassava for the same traits.

The development of fully homozygous genotypes by self-pollination, however, is illusory for hexaploid sweet potato. Even if plants were self-compatible, it would require seven generations of self-pollination to reach an inbreeding coefficient (*f*) of 0.5. The same degree of inbreeding would require just one generation in a diploid species (Grüneberg *et al.*, 2015). As in the case of potato, there are ongoing efforts to introduce some degree of inbreeding in sweet potato through a reduction in the ploidy level (e.g. double-triploids). Plant breeders are of the opinion that outstanding sweet potato clones represent a unique combination of genes rather than unique genes, and epistasis is probably very important (van Rheenen, 1964; Martin, 1987). Similar conclusions have been mentioned for other polyploid species, such as alfalfa (Jones and Bingham, 1995).

Epistasis

Inbreeding research on RTBs (as well as on other crops) is important because it can contribute to understanding of the quantitative genetic foundations of different traits. For example, studies in maize showed a linear relationship between inbreeding depression and *f*, thus leading to the conclusion that epistasis does not seem to play a major role. However, Lamkey and Edwards (1999) emphasized that these studies measured population bulks, and hence examined the average across the whole populations. Evidence of significant epistatic effects could be found if inbreeding depression from lineages of individual genotypes was measured and found to be different. Similarly, the significance of epistasis can be determined if within-family genetic variation can be measured, as shown in the present chapter.

Evidence from molecular biology and field studies clearly shows that genes interact with each other and, when it could be assessed, epistasis is often significant (Lamkey *et al.*, 1995; Wolf and Hallauer, 1997; Lamkey and Edwards, 1999). However, many other studies have failed to detect significant epistatic effects, particularly for complex traits, such as grain yield in maize. Hallauer and Miranda (1988) acknowledged that epistasis played a role in yield and suggested that failures in detecting it were probably the consequence of inadequate genetic models.

Lamkey and Edwards (1999) provided additional insights into the contrasting results regarding the relative importance of epistasis. Generation mean analysis is more likely to find significant epistatic effects than estimates based on the analysis of variance. From a statistical point of view, averages are more powerful than variances. On the other hand, genetic effects in generation mean analysis are fixed and thus the extrapolation of results is limited. Studies based on non-inbred progenitors tend to show a predominance of additive over non-additive genetic effects, whereas the contrary is found when inbred progenitors are used. Moreover, statistical epistasis (the one often found to have limited significance) contrasts markedly with the different views about the so-called biological (also physiological, functional or non-linear molecular) interactions (Cheverud and Routman, 1995; Goodnight, 1999; Moore and Williams, 2005; Álvarez-Castro and Carlborg, 2007; Hansen and Wagner, 2011; Hansen, 2013; Mackay, 2014). Statistical epistasis is defined as deviation from additivity in a mathematical model, whereas the biological perspective envisions epistasis as the interaction among biomolecules within gene regulatory networks and biochemical pathways (Moore and Williams, 2005).

Performance of the best maize hybrids depends mainly on additive and dominance variances but gets an extra boost from epistasis. In other words, what distinguishes the success of the best commercial hybrids from the rest is the extra bit of genetic superiority derived from epistatic effects (Crow, 2000). Similar conclusions have been mentioned by other authors working on maize (Hallauer and Miranda, 1988; Mikel and Dudley, 2006; Melchinger *et al.* 2007) as well as from tree breeding programmes (Zhao *et al.*, 2014). As stated by Goodnight (1999), inbreeding depression and heterosis inevitably require some form of gene interaction (both within and between loci). Non-additive genetic effects, including epistasis and dominance, are key components of heterosis and the superiority of RTB clones identified by breeders and grown by farmers.

The relative importance of additive and non-additive genetic effects assumes a singular relevance in breeding RTBs. As stated above, all genetic effects can be exploited by farmers once an outstanding clone is identified and grown commercially. But breeders face major difficulties capturing the genetic superiority of elite hybrids and transmitting it to the progeny. Results of a large evaluation of cassava in the sub-humid environment of Colombia have recently been published (Ceballos *et al.*, 2016b, 2017a). In that study, more than 30,000 plot data values were used to estimate the breeding value of 108 progenitors used, through 14 years of evaluation. Data on the individual performance of all the clones derived from a given progenitor were also available. Figure 14.2 represents a summary of the results of that study.

Important conclusions could be drawn from the information presented in Fig. 14.2. The response variable is represented by the selection index that integrates four relevant traits of cassava. There is not much variation in breeding values among the evaluated 108 progenitors. There is a large variation in the performance of different progenies from each progenitor, which masks, to a certain extent, the actual breeding values of these progenitors. The best-performing genotypes (two have been marked in Fig. 14.2) are not necessarily linked to progenitors with high and positive breeding value. In fact, there is evidence that an outstanding hybrid can be found in progenitors with negative breeding value. Figure 14.2 illustrates visually the concept that outstanding hybrids are those departing from the average effects defined by breeding values: the extra boost or bit of non-additive genetic superiority mentioned by Crow in 2000. Despite the use of potentially predictive databases and statistical procedures for estimating single-cross performance, maize breeders must still make thousands of crosses to find a few superior commercial hybrids (Troyer, 2006).

Advantages of inbreeding

Inbreeding depression occurs because of the increasing frequency of recessive alleles in the homozygous state in loci that were originally heterozygous and contained favourable dominant alleles affecting the trait. Whereas these dominant alleles usually have individually small, cumulative effects, homozygous deleterious recessive alleles often have relatively large effects on plant vigour. Therefore, selection

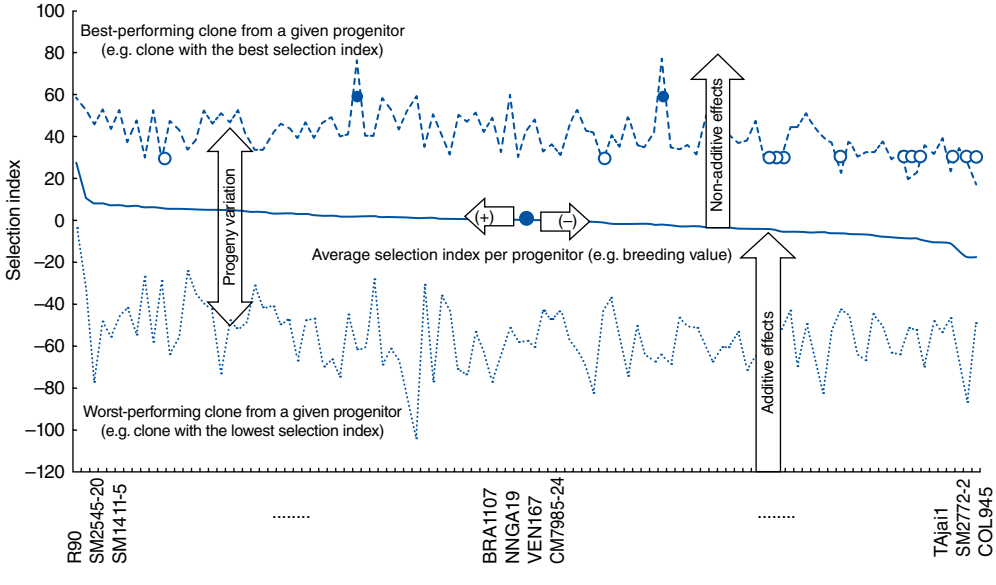


Fig. 14.2. Average selection index (e.g. breeding value) and range of variation for progeny of different progenitors (not all progenitors are named but data from all 108 progenitors are presented). Two clones showed selection index above 60 (black circles). A group of 12 progenitors failed to generate at least one clone with a selection index above 30 (white circles).

against deleterious alleles is thought to be more efficient than selection for favourable additive ones, particularly for traits such as yield (Genter, 1973; Jones and Bingham, 1995). The large genetic load in RTBs (Grüneberg *et al.*, 2015; Ramu *et al.*, 2017; Jansky and Spooner, 2018; Manrique-Carpintero *et al.*, 2018) can be more efficiently reduced through the introduction of inbreeding. However, it should be pointed out that the frequency of homozygosity in polyploid species is extremely low. For an allelic frequency of $p = q = 0.50$, the frequency of homozygous recessive genotypes in a random mating diploid species would be 0.25. The frequencies of the homozygous recessives in tetraploids and hexaploids would be around 0.075 and 0.025, respectively. The expression of a recessively inherited attribute is, therefore, extremely rare in hexaploid sweet potato. Only at high frequencies of the recessive allele ($q > 0.7$) can the desired recessive inherited attribute be observed with detectable frequencies (Grüneberg *et al.*, 2015).

A second important advantage of inbreeding is that it helps to fix favourable genetic combinations. One of the key advantages of using inbred progenitors in breeding commercial hybrids (e.g. maize) is that all genetic effects can be exploited not only by farmers growing the hybrids but also by the breeders (Troyer, 2006). This major advantage justifies the consistent genetic gains observed for maize since the first introduction of hybrids from inbred progenitors almost a century ago. The use of inbred progenitors gradually and consistently fixes favourable gene combinations (e.g. epistasis) in different loci within each progenitor. The simultaneous and coordinated selection in the two inbred progenitors from two heterotic groups guarantees heterozygosity (e.g. dominance) at key loci in the resulting hybrid. In the process of selecting for better performance of the inbred progenitors per se, additive genetic effects are exploited (Duvick, 1999; Troyer, 2006). A clear and frustrating consequence of the use of non-inbred progenitors in the RTB crops is that the favourable gene combinations (painfully assembled in the outstanding clones used as progenitors) are lost as soon as gametes are produced. Consistent gains in complex traits, such as yield, therefore, are difficult to sustain using the standard phenotypic recurrent selection methods.

Inbreeding can help not only reduce genetic load of undesirable traits but also identify useful recessive traits, particularly for diploid RTBs. Other advantages of homozygous progenitors in breeding RTBs include the possibility of implementing the backcross scheme, facilitated germplasm exchange and conservation (as botanical seeds that breed true) and cleaning disease-contaminated planting material of elite hybrids (remaking the hybrid by crossing again with the original progenitors). The possibility of reproducing successful hybrids through botanical seed is one of the appeals of using inbred progenitors in potato. Farmers growing potato have to purchase new batches of planting material because of the high disease pressure, particularly from different viruses. The seed system in this crop requires a continuous supply of tissue-culture-derived material, which is expensive to maintain. Botanical seeds, on the other hand, are usually virus-free. The current view is that botanical seed would be used to produce large seedling plots and from these plots, vegetative tubers could be harvested for commercial planting of potato fields. The use of inbred progenitors would also facilitate the work of molecular research, the identification of quantitative trait loci (QTL) and the implementation of genomic selection. It was because of its enhanced level of homozygosity that genotype AM 560-2 was selected for the sequencing of cassava's genome (Bredeson *et al.*, 2016).

The advantages mentioned above are theoretical. Since limited inbreeding has taken place in RTB crops, there are limited examples of the actual benefits that inbreeding offers for this type of crop. Inbreeding was a key step in the identification of natural (Ceballos *et al.*, 2007) or induced mutations (Ceballos *et al.*, 2008) in cassava, illustrating its relevance for discovering commercially useful recessive mutants. A clear straightforward application of inbreeding is to generate partial inbreds (e.g. S_1 lines) that are homozygous for a monogenic defence trait (for example, resistance to cassava mosaic disease – CMD). The breeding value of S_1 lines that are homozygous for such a trait would double, compared with its S_0 heterozygous progenitor. In fact, inbreeding has been demonstrated to enhance field resistance to cassava brown streak virus (Kaweesi *et al.*, 2016). Japanese sweet potato breeders reported the use of some degree of inbreeding to breed for enhanced dry matter content (Komaki *et al.*, 1998; Lebot, 2010a). It is thought that cassava improvement could greatly benefit from the introduction of inbreeding into

the selection process (Lebot, 2010b). The ongoing efforts to use inbred progenitors in potato are encouraged by the interest in a more efficient exploitation of heterosis in that crop (Gopal, 2014; Jansky and Spooner, 2018).

Assessing Epistasis in Diallel Crosses in RTB Crops

A common feature of all RTBs is their vegetative reproduction. This offers an interesting advantage for genetic studies because individual genotypes within a family can be replicated. This, in turn, allows separating the genetic from the micro-environmental effects. In short, within-family genetic effects can be estimated. Reports in the literature on the relevance of epistasis are not frequent and generally take advantage of the vegetative multiplication that some species offer (Comstock *et al.*, 1958; Stonecypher and McCullough, 1986; Foster and Shaw, 1988; Rönnberg-Wästljung *et al.*, 1994; Rönnberg-Wästljung and Gullberg, 1999; Isik *et al.*, 2003).

A diallel mating design was used to generate F_1 crosses among nine or ten cassava progenitors adapted to three different environments in Colombia. Cassava offers the advantage of having no reported incompatibility barriers; the only limitation for making the required crosses is the need for a synchronized flowering. Inbreeding level of parental lines was considered zero because no self-pollination has been involved in cassava breeding and crosses among related clones are generally avoided. Controlled pollinations were performed following the standard procedures described by Kawano (1980). Many parental clones were initially involved but the parents ultimately used (as well as the number of parents involved) were those that allowed for a balanced set of crosses. Botanical seed were germinated and grown in a screenhouse until the seedlings were 2 months old; at which time they were transplanted to the field. F_1 seedlings were grown in the field for 10 months. Among the many genotypes (>30) from a given F_1 cross, 30 were randomly chosen for these studies based solely on their capacity to produce at least six vegetative cuttings. Minor selection was unavoidable at this stage based on the capacity of the botanical seed to produce vigorous seedlings and plants capable of producing six good quality vegetative cuttings. These factors determined the group of clones representing each F_1 cross in this study. Each of the six stakes was planted in one of three replications at one of two locations.

Statistical model

Analysis of variance was conducted following the expectations for each mean square described in Table 14.2. As is commonly the case, a few plants died or failed to develop normally. Therefore, in a few F_1 crosses fewer than 30 clones were actually evaluated in the field in each of the three replications at the two locations. To take into consideration this lack of uniformity, the harmonic (not the arithmetic) mean was used as *k* (Venkovsky and Barriga, 1992).

The total genetic variance has been partitioned into among-family variation (σ^2_{F1}) and within-family variation ($\sigma^2_{c/F1}$). The among-family variation, in turn, was partitioned into the well-known variances related to general (σ^2_{GCA}) and specific (σ^2_{SCA}) combining ability, which in turn allowed the estimation of σ^2_A and σ^2_D (Griffing 1956; Hallauer and Miranda, 1988):

$$\sigma^2_{GCA} = (\text{Cov.HS}) = 1/4\sigma^2_A + 1/16\sigma^2_{AA} + 1/64\sigma^2_{AAA} + \ldots \text{ etc.} \quad \text{(Eqn 14.1a)}$$

$$\sigma^2_{SCA} = (\text{Cov.FS} - 2\,\text{Cov.HS}) = 1/4\sigma^2_D + 1/8\sigma^2_{AA} + 1/8\sigma^2_{AD} + 1/16\sigma^2_{DD} \ldots \text{ etc.} \quad \text{(Eqn 14.1b)}$$

Genetic parameters were estimated using the following mean squares from Table 14.2:

$$\sigma^2_{GCA} = [MS_{31} - MS_{32} - MS_{41} + MS_{42}]/srk(p-2) \quad \text{(Eqn 14.2a)}$$

$$\sigma^2_{SCA} = [MS_{32} - MS_{42}]/srk \quad \text{(Eqn 14.2b)}$$

Variances for these estimates were computed as follows (Becker, 1985; Vega, 1987; Kang, 1994):

$$\text{Var}(\sigma^2_{GCA}) = \{2/[srk(p-2)]^2\} [(MS^2_{31}/df_{31} + 2) + (MS^2_{32}/df_{32} + 2) + (MS^2_{41}/df_{41} + 2) + (MS^2_{42}/df_{42} + 2)] \quad \text{(Eqn 14.3a)}$$

Table 14.2. Analysis of variance and expected mean squares for a nine-parents diallel design in which the 30 cassava genotypes representing each F_1 cross were clonally propagated.

Source of variation	Degrees of freedom[a]	MS	Expected mean squares
Location (L)	$s - 1$	MS_1	
Rep/L	$s(r - 1)$	MS_2	
F_1	$[p(p - 1)/2] - 1$	MS_3	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{F1*L} + srk\,\sigma^2_{F1}$
GCA	$p - 1$	MS_{31}	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{SCA*L} + rk(p-2)\,\sigma^2_{GCA*L} + srk\,\sigma^2_{SCA} + srk\,(p-2)\,\sigma^2_{GCA}$
SCA	$p(p - 3)/2$	MS_{32}	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{SCA*L} + srk\,\sigma^2_{SCA}$
F_1*L	$(s - 1)([p(p - 1)/2] - 1)$	MS_4	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{F1*L}$
GCA*L	$(s - 1)(p - 1)$	MS_{41}	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{SCA*L} + rk(p-2)\,\sigma^2_{GCA*L}$
SCA*L	$(s - 1)(p(p - 3)/2)$	MS_{42}	$\sigma^2_e + k\,\sigma^2_\varepsilon + rk\,\sigma^2_{SCA*L}$
Error (a)	$s([p(p - 1)/2] - 1)(r - 1)$	MS_5	$\sigma^2_e + k\,\sigma^2_\varepsilon$
Clones/F_1	$(p(p - 1)/2)(k - 1)$	MS_6	$\sigma^2_e + r\,\sigma^2_{c/F1*L} + sr\,\sigma^2_{c/F1}$
Clones/F_1*L	$(p(p - 1)/2)(k - 1)(s - 1)$	MS_7	$\sigma^2_e + r\,\sigma^2_{c/F1*L}$
Error (*b*)	$s(p(p - 1)/2)(k - 1)(r - 1)$	MS_8	σ^2_e

[a] s = number of locations or sites evaluated (2); r = number of replications within each location (3); p = number of parents involved in the diallel crosses (nine or ten); k = number of cloned genotypes representing each F_1 cross (≈ 30).

$$\text{Var}(\sigma^2_{SCA}) = [2/(srk)^2][(MS^2_{32}/df_{32} + 2) + (MS^2_{42}/df_{42} + 2)] \quad \text{(Eqn 14.3b)}$$

In this evaluation, in addition to the usual among-family variation, the vegetative propagation of cassava allowed the analysis of the within-family variation. By cloning individual genotypes, they could be planted in three replications at each of two locations. Therefore, it was possible to partition the within-family variation into its genetic ($\sigma^2_{c/F1}$), genotype by location ($\sigma^2_{c/F1*L}$) and the environmental (σ^2_e) components, as illustrated in Table 14.2.

The within-family analysis allows obtaining information on the relative importance of epistatic effects. In the absence of epistasis, the following equation can be expected (Hallauer and Miranda, 1988):

$$\sigma^2_{c/F1} - 3\,\text{Cov FS} + 4\,\text{Cov HS} \approx 0 \quad \text{(Eqn 14.4)}$$

Occurrence of significant epistatic effects would be demonstrated if the result of the equation is statistically different from zero. The variance for this test is expected to be large because of the complexity of this linear function. The variance was estimated following the principles established in Lynch and Walsh (1998) and Isik *et al.* (2003), as follows:

$$\begin{aligned}&\text{Var (Test Across)}\\ &= \text{Var}\,[\sigma^2_{c/F1} - 3(\sigma^2_{SCA} + 2\sigma^2_{GCA}) + 4\sigma^2_{GCA}]\\ &= \text{Var}\,[\sigma^2_{c/F1} - 3\sigma^2_{SCA} - 6\sigma^2_{GCA} + 4\sigma^2_{GCA}]\\ &= \text{Var}\,[\sigma^2_{c/F1} - 3\sigma^2_{SCA} - 2\sigma^2_{GCA}]\\ &= \text{Var}(\sigma^2_{c/F1}) + \text{Var}(3\sigma^2_{SCA}) + \text{Var}(2\sigma^2_{GCA})\\ &\quad - 6\text{Cov}(\sigma^2_{c/F1}, \sigma^2_{SCA}) - 4\text{Cov}(\sigma^2_{c/F1}, \sigma^2_{GCA})\\ &\quad + 12\,\text{Cov}(\sigma^2_{SCA}, \sigma^2_{GCA})\end{aligned} \quad \text{(Eqn 14.5)}$$

However, since Cov ($\sigma^2_{c/F1}$, σ^2_{SCA}) = 4 Cov ($\sigma^2_{c/F1}$, σ^2_{GCA}) = 0, the formula can be simplified as:

$$\begin{aligned}&\text{Var (Test Across)}\\ &= \text{Var}(\sigma^2_{c/F1}) + 9\,\text{Var}(\sigma^2_{SCA})\\ &\quad + 4\,\text{Var}(\sigma^2_{GCA})\\ &\quad + 12\text{Cov}(\sigma^2_{SCA}, \sigma^2_{GCA})\end{aligned} \quad \text{(Eqn 14.6)}$$

The last term in the equation can be estimated as:

$$\begin{aligned}&\text{Cov}\,(\sigma^2_{SCA}, \sigma^2_{GCA}) = \{(1/srk)\\ &\quad * [1/srk(p-2)]\}\\ &\quad * [\text{Cov}\,(MS_{21}, MS_{31}) - \text{Cov}\,(MS_{21}, MS_{32})\\ &\quad - \text{Cov}\,(MS_{21}, MS_{21}) + \text{Cov}\,(MS_{21}, MS_{22})\\ &\quad - \text{Cov}\,(MS_{22}, MS_{31}) + \text{Cov}\,(MS_{22}, MS_{32})\\ &\quad + \text{Cov}\,(MS_{22}, MS_{21}) - \text{Cov}\,(MS_{22}, MS_{22})]\end{aligned}$$

in the above equation:

$$\begin{aligned}&\text{Cov}\,(MS_{21}, MS_{31}) = \text{Cov}\,(MS_{21}, MS_{32}) = 0\\ &\text{Cov}(MS_{21}, MS_{21}) = \text{Var}(MS_{21})\\ &\text{Cov}(MS_{22}, MS_{22}) = \text{Var}(MS_{22})\end{aligned}$$

Therefore,

$$\begin{aligned}&\text{Cov}(\sigma^2_{SCA},\sigma^2_{GCA})\\&=\{(1/srk)*[1/srk(p-2)]\}*[-\text{Var}(MS_{21})\\&-\text{Var}(MS_{22})+2\text{Cov}(MS_{21},MS_{22})]\\&=-\{2/[s^2r^2k^2(p-2)]\}*\{[(MS_{21})^2/\\&(df+2)]+[(MS_{22})^2/(df+2)]\}\end{aligned}$$

Equation 14.6 can now be written as follows:

$$\begin{aligned}&\text{Var (Test Across)}\\&=\text{Var}(\sigma^2_{c/F1})+9\text{Var}(\sigma^2_{SCA})\\&+4\text{Var}(\sigma^2_{GCA})-12\{2/[s^2r^2k^2(p-2)]\}\\&*\{[(MS_{21})^2/(df+2)]+[(MS_{22})^2/(df+2)]\}\end{aligned}$$

The estimates of additive and dominance variances are overestimated because they contain portions of epistatic variances (Eqns 14.1a and 14.1b).

Following the same principles, standard error of the Epistasis Test for individual location analyses is:

$$\begin{aligned}&\text{Var (Test One Location)}\\&=\text{Var}(\sigma^2_{c/F1})+9\text{Var}(\sigma^2_{SCA})\\&+4\,\text{Var}(\sigma^2_{GCA})-12\,\{2/[r^2k^2s^2(p-2)]\}\\&*[(MS_{22})^2/(df+2)]\end{aligned}$$

Results

The standard among-family diallel analyses of each of the three diallel studies have been published (Calle *et al*., 2005; Jaramillo *et al*., 2005; Cach *et al*., 2006) independently of the within-family genetic analyses (Cach *et al*., 2005; Pérez *et al*., 2005a,b). The standard among-family analyses considered genetic effects to be fixed. In the within-family analyses, however, the main interest was to assess the relative importance of among- and within-family genetic variances and test the significance of epistasis. All effects, therefore, were considered random and normally distributed. The 30 genotypes representing each F_1 cross are clearly a random sample of all possible genotypes that could possibly be derived from the respective parents and contribute to most of the degrees of freedom in the analysis. The only criterion defining which genotype would be used was the capacity to produce six stakes in an environment different from the target environment where the evaluation was conducted. The parents involved in this study were among a group of 25–30 clones and the actual nine or ten parents eventually involved were those that allowed for a balanced set of progenies for the study. Therefore, the main criterion for the selection of the parental lines was their capacity to flower and produce adequate samples of botanical seed from many different crosses. The analysis of variance for the among-family variation follows Griffing's method 4 (Griffing, 1956). Results of these three diallel studies are summarized in Table 14.3 and only for two variables: fresh root yield (FRY) and dry matter content (DMC).

There is a clear contrast in the genetic parameters estimated for the two variables presented in Table 14.3. The estimate of within-family variation is much larger than the among-family counterpart for FRY. In the case of DMC, this is also the case, but differences are not so large. In

Table 14.3. Variance estimates (standard errors within parenthesis) for fresh root yield (FRY) and dry matter content (DMC) in three different diallel sets evaluated in the three environments for cassava production in Colombia. (Ceballos *et al*., 2015.)

Genetic parameter	FRY (t ha^{-1})			DMC (%)		
	Acid soil	Sub-humid	Mid-altitude	Acid soil	Sub-humid	Mid-altitude
σ^2_G (Among)	1.65 (2.95)	13.09 (4.74)	42.78 (13.27)	1.60 (0.66)	0.77 (0.29)	0.35 (0.12)
σ^2_G (Within)	21.08 (2.30)	127.21 (7.65)	288.93 (1918)	3.22 (0.17)	5.56 (0.31)	0.12 (0.12)
σ^2_A	–1.49 (6.32)	17.82 (13.75)	11.88 (24.67)	3.38 (2.40)	1.45 (0.99)	0.99 (0.47)
σ^2_D	9.03 (7.93)	23.87 (11.15)	152.11 (49.08)	0.87 (0.67)	0.77 (0.50)	–0.21 (0.13)
Epistasis test	15.05 (6.74)	100.40 (12.74)	168.91 (39.72)	0.87 (1.29)	4.26 (0.67)	–0.32 (0.92)

fact, for the mid-altitude location, the among-family genetic variation was larger than the within-family component for DMC. Similarly, the non-additive components of genetic variances are much higher (in comparison with the additive variation) in the case of FRY, but that was not the case with DMC.

Interestingly, the analysis of inbreeding depression in a different study (Contreras Rojas *et al.*, 2009) also showed a similar trend. As expected, FRY, which depends more heavily on non-additive genetic effects, is more severely affected by inbreeding depression than DMC. Similar trends have been reported for sweet potato (Komaki *et al.*, 1998). This information is also in agreement with what Fig. 14.2 reveals. The limited variation in breeding values and the confounding effect of the within-family genetic variation for selection index (heavily influenced by FRY) illustrates the challenges for improving FRY further. The information generated by these studies allowed Ceballos and co-workers (2015) to conclude that genomic selection would not be as effective in increasing FRY as it would be for DMC. Empirical data from ongoing research would support these predictions (unpublished data).

References

Álvarez-Castro, J.M. and Carlborg, Ö. (2007) A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167.

Amadi, C.O., Onyeka, J., Chukwu, G.O. and Okoye, B.C. (2015) Hybridization and seed germination of taro (*Colocasia esculenta*) in Nigeria. *Journal of Crop Improvement* 29(1), 106–116, DOI: 10.1080/15427528.2014.980023.

Becker, W.A. (1985) *Manual of Quantitative Genetics*, 4th edn. Academic Enterprises, Pullman, Washington, DC, pp. 95–96.

Balloux, F., Lehmann, L. and de Meeûs T. (2003). The population genetics of clonal and partially clonal diploids. *Genetics* 164, 1635–1644.

Bani-Aameur, F., Lauer, F.I., Veilleux, R.E. and Hilali, A. (1991) Genomic composition of 4x–2x potato hybrids: Influence of *Solanum chocoense*. *Genome* 34, 413–420.

Bousalem, M., Arnau, G., Hochu, I., Arnolin, R., Viader, V., *et al*. (2006) Microsatellite segregation analysis and cytogenetic evidence for tetrasomic inheritance in the American yam *Dioscorea trifida* and a new basic chromosome number in the Dioscoreae. *Theoretical and Applied Genetics* 113, 439–451.

Bradshaw, J.E. (2007) Breeding potato as a major staple crop. In: Kang, M.S. and P.M. Priyadarshan (eds) *Breeding Major Food Staples*. Blackwell Publishing, Ames, Iowa.

Bredeson, J.V., Lyons, J.B., Prochnik, S.E., Wu, G.A., Ha, C.M., *et al*. (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology* 34(5), 562–570. DOI: 10.1038/nbt.3535.

Burton, J.W., Stuber, C.W. and Moll, R.H. (1978) Variability of response to low levels of inbreeding in a population of maize. *Crop Science* 18, 65–68.

Cach N.T., Pérez, J.C., Lenis, J.I., Calle, F., Morante, N., *et al*. (2005) Epistasis in the expression of relevant traits in cassava (*Manihot esculenta* Crantz) for subhumid conditions. *Journal of Heredity* 96, 586–592.

Cach, T.N., Lenis, J.I., Pérez, J.C., Morante, N., Calle, F., *et al*. (2006) Inheritance of relevant traits in cassava (*Manihot esculenta* Crantz) for sub-humid conditions. *Plant Breeding* 125, 177–182.

Calle F., Pérez, J.C., Gaitán, W., Morante, N., Ceballos, H., Llano, G., *et al*. (2005) Diallel inheritance of relevant traits in cassava (*Manihot esculenta* Crantz) adapted to acid-soil savannas. *Euphytica* 144, 177–186.

Ceballos, H., Iglesias C.A., Pérez J.C. and Dixon A.G.O. (2004) Cassava breeding: Opportunities and challenges. *Plant Molecular Biology* 56, 503–515.

Ceballos, H., Sánchez, T., Morante, N., Fregene, M., Dufour, D., *et al*. (2007) Discovery of an amylose-free starch mutant in cassava (*Manihot esculenta* Crantz). *Journal of Agricultural and Food Chemistry* 55(18). 7469–7476.

Ceballos, H., Sánchez, T., Denyer, K., Tofiño, A.P., Rosero, E.A., *et al*. (2008) Induction and identification of a small-granule, high-amylose mutant in cassava (*Manihot esculenta* Crantz). *Journal of Agricultural and Food Chemistry* 56(16), 7215–7222.

Ceballos H, Kawuki, R.S., Gracen, V.E., Yencho G.C. and Hershey, C.H. (2015) Conventional breeding, marker assisted selection, genomic selection and inbreeding in clonally propagated crops: A case study for cassava. *Theoretical and Applied Genetics* 9, 1647–1667.

Ceballos, H., Becerra, L.A., Calle, F., Morante, N., Ovalle, T., *et al*. (2016a) Genetic distance and heterosis in cassava. *Euphytica* 210, 79–92. DOI: 10.1007/s10681-016-1701-7.

Ceballos, H., Pérez, J.C., Orlando, J.B., Lenis, J.I., Morante, N., *et al*. (2016b) Cassava breeding I: The value of breeding value. *Frontiers in Plant Science* 7, 1227. DOI: 10.3389/fpls.2016.01227.

Ceballos, H., Morante N., Calle F., Lenis, J.I. and Salazar, S. (2017a) Cassava breeding. In: Hershey, C. (ed.) *Achieving Sustainable Cultivation of Cassava, Volume 2: Genetics, Breeding*, *Pests and Diseases*, Burleigh Dodds Science Publishing, Cambridge, UK, pp. 49–90.

Ceballos, H., Jaramillo, J.J., Salazar, S., Pineda, L.M., Calle, F., *et al*. (2017b) Induction of flowering in cassava through grafting. *Journal of Plant Breeding and Crop Science* 9, 19–29.

Chaurasiya, P.C., Singh, J., Dikshit, S.N., Mehta, N. (2013) Heterosis, combining ability and gene action studies in sweet potato (*Ipomoea batatas* (L.) Lam. *International Journal of Plant Research* 26(2), 264–270.

Cheverud, J.M. and Routman, E.J. (1995) Epistasis and its contribution to genetic variance components. *Genetics* 139, 1455–1461.

Chipeta, M.M., Bokosi, J.M., Saka, V.W., Benesi, I.R.M. (2013) Combining ability and mode of gene action in cassava for resistance to cassava green mite and cassava mealy bug in Malawi. *Journal of Plant Breeding and Crop Science* 5(9), 195–202.

Chołuj, D., Wisńiewska, A., Szafrański, K.M., Cebula, J., Gozdowski, D., *et al*. (2014). Assessment of the physiological responses to drought in different sugar beet genotypes in connection with their genetic distance. *Journal of Plant Physiology* 171, 1221–1230.

Comstock, R.E., Kelleher, T. and Morrow, E.B. (1958) Genetic variation in an asexual species, the garden strawberry. *Genetics* 43, 634–646.

Contreras Rojas, M, Pérez, J.C., Ceballos, H., Baena, D., Morante, N., *et al*. (2009) Introduction of inbreeding and analysis of inbreeding depression in eight S_1 cassava families. *Crop Science* 49, 543–548.

Cornelius, P.L. and Dudley, J.W. (1974) Effects of inbreeding by selfing and full-sibbing in a maize population. *Crop Science* 14, 815–819.

Crow, J.F. (2000) The rise and fall of overdominance. *Plant Breeding Reviews* 17, 225–257.

Dansi, A., Mignouna, H.D. Pillay, M. and Zok, S. (2001) Ploidy variation in the cultivated yams (*Dioscorea cayenensis–Dioscorea rotundata* complex) from Cameroon as determined by flow cytometry. *Euphytica* 119, 301–307.

Darwin, C. (1876) *The Effects of Cross and Self-fertilisation in the Vegetable Kingdom*, 2nd edn. John Murray, London.

de Freitas, J.P.X., da Silva Santos, V. and de Oliveira, E.J. (2016) Inbreeding depression in cassava for productive traits. *Euphytica* 209,137–145.

de Freitas, J.P.X., Parreira Diniz, R., Alves Santos de Oliveira, S., da Silva Santos, V. and de Oliveira, E.J. (2017) Inbreeding depression for severity caused by leaf diseases in cassava. *Euphytica* 213, 205. DOI: 10.1007/s10681-017-1995-0.

de Jong, H. and P. R. Rowe. 1971. Inbreeding in cultivated diploid potatoes. *Potato Research* 14, 74–83.

de Meeûs, T., Prugnolle, F. and Agnew, P. (2007). Asexual reproduction: Genetics and evolutionary aspects. *Cellular and Molecular Life Sciences* 64(11), 1355–1372. DOI: 10.1007/s00018-007-6515-2.

Dewey, D. R. (1966) Inbreeding depression in diploid, tetraploid, and hexaploid crested wheatgrass. *Crop Science* 6, 144–147.

Duvick, D.N. (1999) Heterosis: Feeding people and protecting natural resources. In: Coors J.G. and Pandey, S. (eds) *Genetics and Exploitation of Heterosis in Crops,* Crop Science Society of America, Inc. Madison, Wisconsin, pp. 19–29.

Easwari Amma, C.S., Sheela, N. and Thankamma Pillai, P.K. (1995) Combining ability analysis in cassava. *Journal of Root Crops* 21(2), 65–71.

Falconer, D.S. (1981) *Introduction to Quantitative Genetics*, 2nd edn. Longman, London and New York.

Foster, G.S., and Shaw, D.V. (1988) Using clonal replicates to explore genetic variation in a perennial plant species. *Theoretical and Applied Genetics* 76, 788–794.

Frascaroli, E., Canè, M.A., Landi, P., Pea, G., Gianfranceschi, L., *et al*. (2007) Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics* 176, 625–644.

Frisch, M., Thiemann, A., Fu, J., Schrag, T.A., Scholten, S., *et al*. (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theoretical and Applied Genetics* 120, 441–450.

Fu, D., Xiao, M., Hayward, A., Fu, Y., Liu, G., *et al*. (2014) Utilization of crop heterosis: A review. *Euphytica* 197, 161–173.

Gamiette, F., Bakry, F. and Ano, G. (1999) Ploidy determination of some yam species (*Dioscorea* spp.) by flow cytometry and conventional chromosomes counting. *Genetic Resources and Crop Evolution* 46, 19–27.

Genter, C.F. (1973) Comparison of SI and testcross evaluation after two cycles of recurrent selection in maize. *Crop Science* 13, 524–527.

Goodnight, C.J. (1999) Epistasis and heterosis. In: Coors J.G. and Pandey, S. (eds) *Genetics and Exploitation of Heterosis in Crops*. Crop Science Society of America, Inc. Madison, Wisconsin, pp. 59–68.

Goff, S.A. (2011) A unifying theory for general multigenic heterosis: Energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytologist* 189, 923–937. DOI: 10.1111/j.1469-8137.2010.03574.x.

Gopal, S., (2014) Heterosis breeding in potato. *Agricultural Research* 3(3), 204–217.

Griffing, B. (1956) Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Sciences* 9, 463–493.

Groose, R.W., Talber, L.E., Kojis, W.P. and Bingham, E.T. (1989) Progressive heterosis in autotetraploid alfalfa: Studies using two types of inbreds. *Crop Science* 29, 1173–1177.

Grüneberg, W.J., Ma, D., Mwanga, R.O.M., Carey, E.E., Huamani, K., *et al*. (2015) Advances in sweetpotato breeding from 1992 to 2012. In: Low, J., Nyongesa, M., Quinn, S. and Parker, M. (eds) *Potato and Sweetpotato in Africa: Transforming the Value Chains for Food and Nutrition Security*. CAB International, Wallingford, UK, pp. 3–68.

Gurmu, F., Shimelis Hussein, S. and Laing, M. (2018) Combining ability, heterosis, and heritability of storage root dry matter, beta-carotene, and yield-related traits in sweetpotato. *Hortscience* 53(2), 167–175.

Hallauer, A.R. and Miranda Fo, J.B. (1988). *Quantitative Genetics in Maize Breeding*, 2nd edn. Iowa State University Press, Ames, Iowa.

Hallauer, A.R. and Sears, J.H. (1973) Changes in quantitative traits associated with inbreeding in a synthetic variety of maize. *Crop Science* 13, 327–330.

Hansen, T.F. (2013) Why epistasis is important for selection and adaptation. *Evolution* 67, 3501–3511.

Hansen, T.F. and Wagner G.P. (2011) Modeling genetic architecture: A multilinear theory of gene interaction. *Theoretical Population Biology* 59, 61–86.

Iramu, E., Wagih, M.E. and Singh, D. (2010) Genetic hybridization among genotypes of Taro (*Colocasia esculenta*) and recurrent selection for leaf blight resistance. *Indian Journal of Science and Technology* 3(1), 96–101. DOI: 10.17485/ijst/2010/v3i1/29656.

Isik, F, Li, B., and Frampton, J. (2003) Estimates of additive, dominance and epistatic genetic variances from a clonally replicated test of loblolly pine. *Forest Science* 49(1), 77–88.

Ivancic, A. and Lebot, V. (2000) *The Genetics and Breeding of Taro*. Séries Repères, CIRAD, Montpellier, France.

Jansky, S.H. and Spooner, D.M. (2018) The evolution of potato breeding. *Plant Breeding Reviews* 41, 169–214.

Jaramillo, G., Morante, N., Pérez, J.C., Calle, F., Ceballos, H., *et al*. (2005) Diallel analysis in cassava adapted to the midaltitude valleys environment. *Crop Science* 45, 1058–1063.

Jennings, D.L. and C.A. Iglesias, C.A. (2002) Breeding for crop improvement. In: Hillocks, R.J., Thresh J.M. and Belloti, A.C. (eds) *Cassava: Biology, Production, and Utilization*. CAB International, Wallingford, UK, pp. 149–146.

Jones, J.S. and Bingham, E.T. (1995) Inbreeding depression in alfalfa and cross-pollinated crops. *Plant Breeding Reviews* 13, 209–233.

Kang, M.S. (1994) *Applied Quantitative Genetics*. Louisiana State University, Baton Rouge, Louisiana, 157 pp.

Kawano, K. (1980) Cassava. In: Fehr, W.R. and Hadley, H.H. (eds) *Hybridization of Crop Plants*. ASA, CSSA, Madison, Wisconsin, pp. 225–233.

Kaweesi, T., Kyaligonza, V., Baguma Y., Kawuki, R. and Ferguson, M. (2016) Inbreeding enhances field resistance to cassava brown streak viruses. *Journal of Plant Breeding and Crop Science* 8(8), 138–149.

Kawuki, R.S., Nuwamanya E., Labuschagne, M.T., Herselman, L. and Ferguson, M. (2011) Segregation of selected agronomic traits in six S_1 cassava families. *Journal of Plant Breeding and Crop Science* 3(8), 154–160.

Komaki, K., Katayama, K., and Tamiya, S. (1998) Advancement of sweet potato breeding for high starch content in Japan. *Tropical Agriculture* 75(1/2), 220–223.

Lamkey, K.R. and Edwards, J.W. (1999) Quantitative genetics of heterosis. In: Coors J.G. and Pandey, S. (eds) *Genetics and Exploitation of Heterosis in Crops*. Crop Science Society of America, Inc., Madison, Wisconsin, pp. 31–48.

Lamkey, K.R., Schnicker, B.J. and Melchinger, A.E. (1995) Epistasis in an elite maize hybrid and choice of generation for inbred line development. *Crop Science* 35, 1272–1281.

Lebot, V. (2010a) Sweet potato. In: Bradshaw, J.E. (ed.) *Root and Tuber Crops*. Springer, New York, pp. 97–125.

Lebot V. (2010b) Tropical root and tuber crops. In: Verheye, W.H. (ed.) *Soils, Plant Growth and Crop Production,* Volume 2. EOLSS UNESCO Online Encyclopaedia of life support systems, Oxford, UK.

Lin, K.-H., Lai, Y.-C., Chang, K.-Y., Chen, Y.-F., Hwang, S.-Y., *et al*. (2007) Improving breeding efficiency for quality and yield of sweet potato. *Botanical Studies* 48, 283–292.

Lindhout, P., De Vries, M., Ter Maat, M., Ying, S., Viquez-Zamora, M., *et al*. (2018). Hybrid potato breeding for improved varieties. In: Wang-Pruski, G. (ed.) *Achieving Sustainable Cultivation of Potatoes*, Volume 1. Burleigh Dodds Science Publishing, Cambridge, UK, pp. 99–124.

Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, Massachusetts, pp. 558–563 (chapter 18) and pp. 813–816 (appendix 1).

Mackay, T.F.C. (2014) Epistasis and quantitative traits: Using model organisms to study gene–gene interactions. *Nature Reviews Genetics* 15, 22–33. DOI: 10.1038/nrg3627.

Manrique-Carpintero, N.C., Coombs, J.J., Pham, G.M., Parker, F., Laimbeer, *et al*. (2018) Genome reduction in tetraploid potato reveals genetic load, haplotype variation, and loci associated with agronomic traits. *Frontiers in Plant Science* 9, 944. DOI: 10.3389/fpls.2018.00944.

Martin, F.W. (1987) Preservation of sweet potato germplasm as population. In: *Exploration, Maintenance, and Utilization of Sweet Potato Genetic Resources. Report of the First Sweet Potato Planning Conference*. International Potato Center (CIP), Lima, Peru, pp. 159–167.

McDavid, C.R. and Alamu, S. (1976) Promotion of flowering in tannia (*Xanthosoma sagittifolium*) by gibberellic acid. *Tropical Agriculture (Trinidad)* 53(4), 373–374.

Melchinger, A.E., Utz, H.F., Piepho, H.-P., Zeng Z.-B. and Schön, C.C. (2007) The role of epistasis in the manifestation of heterosis: A systems-oriented approach. *Genetics* 177, 1815–1825.

Mendoza, H.A. and Haynes, F.L. (1974) Genetic basis of heterosis for yield in the autotetraploid potato. *Theoretical and Applied Genetics* 45, 21–25.

Mikel, M.A. and Dudley, J.W. (2006) Evolution of North American dent corn from public to proprietary germplasm. *Crop Science* 46, 1193–1205.

Moore, J.H. and Williams, S.M. (2005) Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *Bioessays* 27(6), 637–46.

Mwanga, R.O.M., Andrade, M.I., Carey, E.E., Low, J.W., Yencho, G.C., *et al*. (2017). Sweetpotato (*Ipomoea batatas* L.). In: Campos, H. and Caligari, P.D.S. (eds) *Genetic Improvement of Tropical Crops*. Springer, Cham, Switzerland.

Nassar, N.M.A., Graciano-Ribeiro, D., Gomes P.F. and Hashimoto, D.Y.C. (2010) Alterations of reproduction system in a polyploidized cassava interspecific hybrid. *Hereditas* 147, 58–61.

Obidiegwu, J.E., Kendabie, P., Obidiegwu, O. and Amadi, C. (2016) Towards an enhanced breeding in cocoyam: A review of past and future research Perspectives. *Research & Reviews: Journal of Botanical Sciences* 5(4), 22–33.

Ortiz, R. (1997) Secondary polyploids, heterosis, and evolutionary crop breeding for further improvement of the plantain and banana (*Musa* spp. L) genome. *Theoretical and Applied Genetics* 94, 1113–1220.

Oumar, D., Sama, A.E., Adiobo, A. and Zok, S. (2011) Determination of ploidy level by flow cytometry and autopolyploid induction in cocoyam (*Xanthosoma sagittifolium*). *African Journal of Biotechnology* 10(73), 16491–16494.

Pérez, J.C., Ceballos, H., Calle, F., Morante, N., Gaitán, W., *et al*. (2005a) Within-family genetic variation and epistasis in cassava (*Manihot esculenta* Crantz) adapted to the acid-soils environment. *Euphytica* 145, 77–85.

Pérez, J.C., Ceballos, H., Jaramillo, G., Morante, N., Calle, F., *et al*. (2005b) Epistasis in cassava adapted to mid-altitude valley environments. *Crop Science* 45, 1491–1496.

Pineda, L.M., Morante, N., Salazar, S., Hyde, P., Setter, T., *et al*. (2018a) Induction of flowering I: Photoperiod extension through a red lights district. *IVth GCP21 International Cassava Conference*, Cotonou, Benin, 10–15 June 2018.

Pineda, L.M., Morante, N., Salazar, S., Hyde, P., Setter, T, *et al*. (2018b) Induction of flowering II: Night breaks as an alternative for photoperiod extension. *IVth GCP21 International Cassava Conference*, Cotonou, Benin, 10–15 June 2018.

Pineda, L.M., Hyde, P., Setter, T., Morante, N., Salazar, S., *et al*. (2018c) Induction of flowering III: The potential of plant growth regulators. *IVth GCP21 International Cassava Conference*, Cotonou, Benin, 10–15 June 2018.

Pineda, L.M., Yu, B., Yinong, T., Morante, N., Salazar, S., *et al*. (2018d) Induction of flowering IV: The potential of pruning young branches. IVth GCP21 International Cassava Conference, Cotonou, Benin, 10–15 June 2018.

Quero-García, J., Letourmy, P., Ivancic, A., Feldmann, P., Courtois, B., *et al*. (2009) Hybrid performance in taro (*Colocasia esculenta*) in relation to genetic dissimilarity of parents. *Theoretical and Applied Genetics* 119, 213–221.

Ramu, P., Esuma, W., Kawuki, R., Rabbi I.Y., Egesi, C., *et al*. (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics* 49, 959–963. DOI: 10.1038/ng.3845.

Robinson, H.F. and Cockerham, C.C. (1961) Heterosis and inbreeding depression in populations involving two open-pollinated varieties of maize. *Crop Science* 1, 68–71.

Rönnberg-Wästljung, A.C. and Gullberg, U. (1999) Genetics of breeding characters with possible effects on biomass production in *Salix viminalis* (L.). *Theoretical and Applied Genetics* 98, 531–540.

Rönnberg-Wästljung, A., Gullberg, U. and Nilsson, C. (1994) Genetic parameters of growth characters in *Salix viminalis* grown in Sweden. *Canadian Journal of Forest Research* 24, 1960–1969.

Sanford, J.C. and Hanneman Jr, R.E. (1982) A possible heterotic threshold in the potato and its implications in breeding. *Theoretical and Applied Genetics* 61, 151–159.

Sprague, G.F. (1983) Heterosis in maize: Theory and practice. *Theoretical and Applied Genetics* 6, 47–70.

Stonecypher, R.W. and McCullough, R.B. (1986) Estimates of additive and non-additive genetic variance from a clonal diallel of Douglas-fir *Pseudotsuga menziesii* (Mirb.) Franco. In: *IUFRO Conference Proceedings: A Joint Meeting of Working Parties on Breeding Theory, Progeny Testing, Seed Orchards,* Williamsburg, Virginia. North Carolina State University, Raleigh, North Carolina, pp. 211–227.

Stuber, C.W. (1994) Heterosis in plant breeding. *Plant Breeding Reviews* 12, 227–251.

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjarisn, T. and Lander, E.S. (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* 132, 823–839.

Tan, S.L., Nakatani, M. and Komaki, K. (2007) Breeding of sweetpotato. In: Kang, M.S. and Priyadarshan, P.M. (eds) *Breeding Major Food Staples*. Blackwell Publishing, Ames, Iowa, pp. 333–363.

Troyer, A.F. (2006) Adaptedness and heterosis in corn and mule hybrids. *Crop Science* 46, 528–543.

van Rheenen, H.A. (1964) Breeding research in sweet potato, *Ipomoea batatas* Poir. II. Selfing and crossing of sweet potato clones. Pre-treatment of sweet potato seed. *Euphytica* 13, 94–99.

Vega-O., P.C. (1987) *Introducción a la Teoría de Genética Cuantitativa*. Universidad Central de Venezuela Press, Caracas, Venezuela.

Vencovsky R. and Barriga, P. (1992) *Genética Biométrica no Fitomelhoramento*. Sociedade Brasileira de Genética, Ribeirão Preto, Brazil. 486 pp.

Volin, R.B. and Beale, A.J. (1981) Genetic variation in cocoyam (*Xanthosoma sp*.) hybrids. *Proceedings of the Florida State Horticultural Society* 94, 235–238.

Wang, C., Lentini, Z., Tabares, E., Quintero, M., Ceballos, H., *et al*. (2011). Microsporogenesis and pollen formation in cassava (*Manihot esculenta* Crantz). *Biologia Plantarum* 55, 469–478.

Wilson, J.E. (1990) *Taro Breeding. Agro-Facts*. IRETA Publication No. 3/89. Apia, Western Samoa.

Wolf, D.P. and Hallauer, A.R. (1997) Triple testcross analysis to detect epistasis in maize. *Crop Science* 37, 763–770.

Wolfe, M., Bauchet, G.J., Chan, A.W., Lozano, R., Ramu, P., *et al*. (2019). Introgressed *Manihot glaziovii* alleles in modern cassava germplasm benefit important traits and are under balancing selection. *bioRxiv*. DOI: 10.1101/624114.

Wright, S. (1922) The effects of inbreeding and crossbreeding on guinea pigs. USDA Bul. 1121. USDA, Washington DC.

Zhang, Y.D., Fan, X.M., Yao, W.H., Piepho, H.-P., Kang M.S. (2016) Diallel analysis of four maize traits and a modified heterosis hypothesis. *Crop Science* 56, 1115–1126.

Zhao, X., Bian, X., Liu, M., Li, Z., Li, Y., *et al*. (2014) Analysis of genetic effects on a complete diallel cross test of *Betula platyphylla*. *Euphytica* 200, 221–229.

15 Genomic Selection in Rice: Empirical Results and Implications for Breeding

Nourollah Ahmadi[1,2]*, Jérôme Bartholomé[1,2], Tuong-Vi Cao[1,2] and Cécile Grenier[1,2]
[1]*CIRAD, UMR AGAP, Montpellier, France;* [2]*AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France*

Introduction

Genomic selection (GS) has arisen from the conjunction of new high-throughput marker technologies and new statistical methods that allow the analysis of the genetic architecture of complex traits in the framework of infinitesimal model effects, instead of a model of limited numbers of quantitative trait loci (QTL) of varying effects. It refers to methods that use genome-wide dense markers, mainly single nucleotide polymorphisms (SNPs), for the prediction of genetic values with enough accuracy to allow selection on that prediction alone. It consists of (i) using all markers (often large numbers) simultaneously to build a model of genotype–phenotype relationships in a training population (TP), thus accounting also for linkage disequilibrium (LD) among markers, and (ii) using the model to predict the genomic estimate of breeding values (GEBV) of candidates in a breeding population (CP) (Meuwissen *et al.*, 2001; Heffner *et al.*, 2009). It extends the use of markers to breeding highly polygenic traits, such as yield, tolerance to abiotic stresses and resource-use efficiency.

The effectiveness of GS depends, among other factors, on the degree of correlation between the predicted GEBV and the true genetic value, i.e. the predictive ability of prediction (PA). In practice, PA of genomic prediction is evaluated by the correlation between GEBV and the realized phenotype.

Prospects for the applications of GS in plant breeding have given rise to many studies using a simulation approach or empirical data to analyse the effects of factors that affect the PA of genomic predictions. These interrelated factors include the characteristics of TP and CP and relationship between the two populations, the characteristics of the target phenotypic trait, the characteristics of genotypic data (marker density, LD and minor allele frequency) and the prediction methods.

Characteristics of the TP that affect PA of genomic prediction include its size, its structure and its relatedness with the CP. Meuwissen (2009) showed that the size of TP depended on the effective size of the population (Ne) and the length of the genome (L) in Morgans (M). Likewise, predicting the breeding values of unrelated individuals required much larger TP than predicting individuals that are progeny of the TP. In the two cases, the optimal size of the TP would be $2 \times Ne \times L$. Simulation work of Meuwissen (2009) also showed that the optimal number of markers to predict breeding values of unrelated individuals would be $10 \times Ne \times L$. In the case of rice ($Ne = 50$, $L = 15$ M), these findings would

* Email: nourollah.ahmadi@cirad.fr

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

imply a TP of 1500 individuals genotyped with 7500 markers, provided the markers are evenly distributed across the genome. The characteristics of the target trait reported to influence the PA of genomic predictions include its heritability, the number of QTL, the distribution of their allelic effects and frequencies, and the relative magnitude of additive and non-additive genetic variances (Hayes *et al*., 2009; Jannink *et al*., 2010; Howard *et al*., 2014). Characteristics of genotypic data include marker density and distribution along the genome, the extent of LD and the minor allele frequency (MAF). The accuracy of different prediction methods depends on the above listed factors, i.e. the characteristics of the target trait, the density and distribution of the markers, the size and the structure of the TP, and the degree of relatedness between TP and CP (Heslot *et al*., 2012; de los Campos *et al*., 2015; Crossa *et al*., 2017).

In order for GS to become a practical method for plant breeding, especially for major annual crops, at least three methodological issues need to be further addressed: (i) method for the establishment of the TP for making selection decisions in pedigree breeding within the progeny of biparental crosses; (ii) method to account for information available on genes/QTL involved in the determinism of complex traits; and (iii) method to account for genotype-by-environment interactions (GEI) as observed in multilocation trials of advanced breeding lines and/or in managed-environment experiments to assess tolerance to abiotic stresses (drought, extreme temperatures, salinity, etc.).

Here we present: (i) a review of empirical studies analysing factors that affect the PA of GEBV in rice, and (ii) some of the results of analysis of the above-mentioned issues (TP for pedigree breeding, accounting for GEI and trait-specific markers) by CIRAD's Genetic and Varietal Innovation team involved in the implementation of different rice breeding programmes worldwide.

Factors Affecting the Predictive Ability of Genomic Prediction in Rice

The empirical studies analysing factors that affect the predictive ability of genomic prediction in rice are listed in Table 15.1. In essence, individual empirical studies are not the most powerful tool for the analysis of effects of the inter-related factors that affect the predictive ability of genomic prediction. However, when congruent, the results of such studies provide valuable practical indications.

Characteristics of the training population

Characteristics of the TP that affect the PA of genomic predictions include its size, its structure and its relatedness with the CP. In the prediction experiments based on cross validation that we reviewed, the size of the TP was around four-fifths of the total number of entries available, which varied from 110 to 575. The resulting size of the TP was much below the theoretical value of 1500 for an *Ne* of 50. Indeed, as reported by Grenier *et al*. (2015), the *Ne* higher than 50 can be observed in a population of rather limited genetic base. These small sizes of the TP are probably one of the causes of the low PA of genomic prediction observed in all studies, even for phenotypic traits of high heritability, such as days to flowering (Table 15.1). The effect of population structure has been analysed in several studies. Using a diversity panel of 413 accessions composed of representatives of the *Oryza sativa* major genetic groups (*indica*, temperate *japonica*, tropical *japonica*, *aus*), Guo *et al*. (2014) analysed the effect of population structure. They reported that the most accurate predictions were obtained by stratified sampling of the training set, i.e. presence of representatives of each genetic group in both training and validation sets. Using the same population, Isidor *et al*. (2015) compared the PA of five algorithms of optimization of the TP (stratified sampling, mean of the coefficient of determination [CDmean], mean of predictor error variance [PEVmean], stratified CDmean [StratCDmean] and random sampling). In the presence of strong population structure, the stratified sampling showed the highest PA for all traits. Grenier *et al*. (2015) reached similar conclusions with breeding lines extracted from four synthetic populations of tropical *japonica*. Whatever the trait, identical PAs were obtained when the lines composing TP and CP were randomly sampled (without

Table 15.1. Genomic prediction studies conducted on rice.

Plant material	Phenotypic data[a]	Genotypic data[b]	Type of prediction experiment	Statistical methods[c]	Range of accuracy of GEBV[d]	Main conclusion[e]	Reference
Highly structured diversity panel of 413 accessions	15 traits of rather high heritability, including DTF, PH and protein content	36,901 SNPs (1 SNP per 10 Kb)	Cross validation	GBLUP, GBLUP-CPS	DTF: 0.44–0.66 PH: 0.50–0.75	Prediction accuracy was affected by the genomic relationship between TP and CP and by genomic heritability in the TP and CP.	Guo *et al*. (2014)
	8 traits including DTF, PH and GY	36,901 SNPs (1 SNP per 10 Kb)	Cross validation	GBLUP	DTF: 0.25–0.60 PH: 0.25–0.55 GY: 0.20–0.50	Maximizing the phenotypic variance captured by the training set is important for optimal performance. Stratified sampling of the training set ensures better accuracy than sampling based on the CDmean.	Isidro *et al*. (2015)
110 Asian cultivars	8 traits including DTF	3,071 SNPs	Cross validation	rr-BLUP, ENet, GBLUP, RKHS, RF, Lasso, BL, EBL, wBSR	DTF: 0.65–0.85	Reliability depended to a great extent on the targeted traits. Reliability was low when only a small number of cultivars were used for validation.	Onogi *et al*. (2015)
369 Elite breeding lines	6 traits including DTF and GY	73,147 SNPs	Cross validation	rr-BLUP, BL, RKHS, RF,	DTF: 0.35–0.65 PH: 0.15–0.35 GY: 0.10–0.30	Using one marker every 0.2 cM was sufficient for genomic selection in this collection of rice breeding material. rr-BLUP was the most efficient statistical method for GY where no marked effect of QTL was detected by GWAS.	Spindel *et al*. (2015)
343 S2:4 lines extracted from a synthetic population	DTF, GY and PH	8,336 SNPs 1 marker per 44.8 kb	Cross validation	rr-BLUP, GBLUP, Lasso, BL	DTF: 0.20–0.30 PH: 0.50–0.60 GY: 0.20–0.31	Accuracy of GEBV was affected by (i) relatedness between TP and CP and (ii) trait heritability and interaction between traits and all the other factors studied (prediction models, LD, MAF, composition of the TP).	Grenier *et al*. (2015)

Continued

Table 15.1. Continued.

Plant material	Phenotypic data[a]	Genotypic data[b]	Type of prediction experiment	Statistical methods[c]	Range of accuracy of GEBV[d]	Main conclusion[e]	Reference
575 F1 hybrids	8 traits including GY and PH	2,395,866 SNPs	Cross validation	GBLUP, GBLUP dominance effects	PH: 0.45–0.86 GY: 0.13–0.34	Model including the dominance effect provided more accurate prediction, particularly in multi-traits scenario for a low-heritability target trait, with highly correlated auxiliary traits.	Wang *et al*. (2017)
Diversity panel of 284 accessions + 97 elite lines derived from crosses between 31 accessions of the panel	DTF, NI and PW	43,686 SNPs	Progeny prediction	BayesB, GBLUP, RKHS	DTF: 0.58–0.65 PW: 0.55–0.62	The diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel.	Ben Hassen *et al*. (2017)
			Progeny prediction and multi-environment prediction	Multi-environment models, GBLUP and RKHS	DTF: 0.70–0.92 PW: 0.55–0.85	Genomic prediction accounting for G×E interactions offers an effective framework for breeding simultaneously for adaptation to an abiotic stress and performance under normal cropping conditions in rice.	Ben Hassen *et al*. (2018)
Diversity panel of 280 accessions	DTF, GY and PH	250,000 SNPs	Cross validation and multi-environment prediction	Multi-environment models, GBLUP and RKHS	DTF: 0.60–0.93 GY: 0.40–0.85 PH: 0.55–0.85	Selection of trait-specific markers and multi-environment models improve genomic predictive ability in rice.	Bhandari *et al*. (2019)
Diversity panel of 225 accessions + 95 elite lines	Arsenic content in the flag-leaf (FL-As) and in the cargo grain (CG-As)	22,370 SNPs	Across populations	BayesB, GBLUP and RKHS	FL-As: 0.35–0.45 CG-As: 0.45–0.55	Genomic prediction offers the most effective marker assisted breeding approach for ability to prevent arsenic accumulation in rice grains.	Frouin *et al*. (2019)

[a]DTF: days to flowering; PH: plant height; GY: grain yield; NI: nitrogen index; and PW: panicle weight.
[b]SNP: single nucleotide polymorphism.
[c]GBLUP: genomic best linear unbiased prediction; rr-BLUP: ridge regression best linear unbiased prediction; ENet: elastic net; RKHS: reproducing kernel Hilbert space regression; RF: Random forest; BL: Bayesian lasso; EBL: extended Bayesian lasso; and wBSR: weighted Bayesian shrinkage regression.
[d]GEBV: genomic estimate of breeding value.
[e]TP: training population; CP: candidates in a breeding population; QTL: quantitative trait locus/loci; GWAS: genome-wide association studies; LD: linkage disequilibrium; and MAF: minor allele frequency.

considering their membership in one of the four subpopulations) or when TP and CP were composed of a balanced share of each of the four subpopulations. On the other hand, significantly lower PA was observed when the TP comprised all the lines of three of the subpopulations and the CP comprised all the lines of the fourth subpopulation.

The importance of relatedness between the TP and the CP was further confirmed by Ben Hassen *et al.* (2017) in their progeny prediction experiment. They showed that the size of the TP (284 accessions including 31 accessions that were the parents of CP) could be reduced to one-third without a significant decrease in PA, if the accessions making up this one-third were the most related to the accessions of the CP and included the 31 parents.

Effect of trait characteristics

In most rice genomic prediction experiments, the PAs for traits, such as days to flowering (DTF) and plant height (PH), are higher than that for grain yield (GY) or its proxy, panicle weight (Table 15.1). Almost all those experiments also report much higher heritability for DTF and PH than for GY, confirming the positive relationship between trait heritability and the PA of genomic prediction. For instance, Grenier *et al.* (2015) reported highly significant differences for PA between PH and GY, in their population of 343 S2:4 lines extracted from a synthetic *japonica* population. In their experiment, the heritability of PH and GY was 0.58 and 0.29, respectively. Likewise, Guo *et al.* (2014), using a diversity panel of 413 accessions, reported variation for PA ranging from 0.44 to 0.84 according to phenotypic traits and attributed the variation to differences in traits heritability. Less straightforward indications are available regarding the effect of the genetic architecture of the target trait. Using genotypic and phenotypic data from 300 elite *indica* lines, Spindel *et al.* (2015) reported lower PA of genomic predictions for GY compared with DTF. When the same data served for QTL mapping by GWAS, no large-effect QTL were detected for GY, whereas for DTF a single very large-effect QTL was detected.

Effect of characteristics of genotypic data (marker density, linkage disequilibrium and minor allele frequency)

The number of markers used in the rice genomic predictions experiments varied from 3071 to 2,395,866 (Table 15.1), representing densities of 8 to 560 SNPs per Mb, often much higher than the theoretical number of markers, 7500 (density of 19.5 SNPs per Mb). Analysis of the effect of marker density on the PA of genomic prediction by Spindel *et al.* (2015), using different subsets of the 73,147 SNPs available, showed that number of markers >7500 did not improve the PA of genomic predictions among a population of 363 elite *indica* lines. Grenier *et al.* (2015) reported that the largest PAs for PH and GY, among the 343 S2:4 tropical *japonica* lines, were achieved with a marker density of 13 SNPs per Mb. Bhandari *et al.* (2019) compared the PA of genomic prediction for GY, PH and DTF in a reference population of the International Rice Research Institute's (IRRI's) rainfed lowland breeding programme, using 15 incidence matrices, with the number of markers ranging from 3000 to 215,000 SNPs. The 15 matrices were established by combining five thresholds of LD ($r^2 \leq 0.25$, $r^2 \leq 0.50$, $r^2 \leq 0.75$, $r^2 \leq 0.90$ and $r^2 \leq 1$) with three thresholds of MAF (≥2%, ≥5% and ≥25%). Significant differences for PA were observed only for LD variation. Whatever the trait, the incidence matrices with LD value of $r^2 \leq 0.50$ led to a PA that was significantly higher than the one with $r^2 > 0.5$. It was concluded that a marker density of 27 SNPs per Mb was sufficient.

These variations of the optimal density of markers between different studies are most probably attributable to differences in the extent of LD within the plant material used in each study. The low optimal density of markers reported by Grenier *et al.* (2015) is probably attributable to the very large extent of LD ($r^2 = 0.59$ at pairwise distance between markers of 0–25 kb and $r^2 = 0.2$ at distance of 0.9–1.5 Mb) among the S3:4 lines belonging to the tropical *japonica* genetic group, known for its large LD. The much higher optimal density of markers reported by Bhandari *et al.* (2019) should be attributed to the rather low extent of LD ($r^2 = 0.103$ at pairwise distances between 0 and 25 kb) among the diversity panel composed of accessions belonging to the *indica* and *aus* genetic groups.

Effect of prediction methods

At least ten different methods were used across all studies. Genomic best linear unbiased prediction (GBLUP) was the most used method. Studies generally compared two to four methods, but Onogi *et al*. (2015) and Grenier *et al*. (2015) compared up to nine methods. Using eight phenological and morphological traits of 110 rice cultivars, mainly developed in Japan, Onogi *et al*. (2015) compared the performances of nine genomic prediction methods: GBLUP, reproducing kernel Hilbert space regression (RKHS), Lasso, elastic net, random forest, Bayesian lasso, extended Bayesian lasso, weighted Bayesian shrinkage regression and the average of all methods. GBLUP was the most accurate for one trait, RKHS and the average of all methods for two traits, and random forest for three traits. The methods were also compared through simulation. Conditions considered in the simulation experiments included factors related to traits (number of QTL and heritability) and to TP (size and extent of LD). The Bayesian lasso, the extended Bayesian lasso and the averaging methods showed stable performance across the simulated scenarios, whereas the other methods, except weighted Bayesian shrinkage regression (which performed poorly in most scenarios), had specific areas of applicability. Similar interactions between prediction methods and phenotypic traits were reported by Grenier *et al*. (2015) and Spindel *et al*. (2015).

Designing Training Population for Pedigree Breeding

An important question in the application of GS in the context of pedigree breeding among the progeny of biparental crosses is how the TP should be constructed to predict progeny from individual crosses. Early attempts to answer this question relative to maize breeding have explored mainly options where the TP was constructed from multiple related or unrelated small biparental families. For instance, the TP was composed of full-sib doubled-haploid (DH) lines that formed the CP (Goddard and Hayes, 2009; Lorenzana and Bernardo, 2009), or half-sib DH lines of the CP, or combinations of full-sib, half-sib and non-related lines (Meuwissen, 2009; Riedelsheimer *et al*., 2013). More recently, a second approach was investigated in a number of crops. It consists of using a reference set to train the prediction model and in using this model to predict the performances of progenies from biparental crosses between members of the reference set. For instance, Hofheinz *et al*. (2012) used a reference set of 310 inbred sugar beet lines to predict the test cross value of 56 inbred progenies derived from eight crosses between six lines of the reference set, and reported average prediction accuracy of 0.79 for sugar content. Sallam *et al*. (2015) used a training set of 168 barley lines and five sets of 96 progeny lines representative of the breeding lines developed in five consecutive years (the training set included the parents of the progeny sets) and reported a prediction accuracy of around 0.50 for grain yield. Likewise, Gezan *et al*. (2017) used a panel representative of the University of Florida's strawberry breeding programme and sets of progenies derived from the circular mating of 31 members of the panel and reported a prediction accuracy ranging from 0.16 to 0.77, depending on the traits and model fitting method used.

In rice, the first empirical evaluation of performances of GS for pedigree breeding in the progeny of biparental crosses was reported by our team (Ben Hassen *et al*., 2017). The TP was represented by the reference population of the Consiglio Ricerche in Agricoltura (CRA; Vercelli, Italy) rice breeding programme composed of 284 accessions belonging to the *japonica* group, and the CP was composed of 97 advanced (F_5–F_7) inbred lines derived from 36 biparental crosses involving 31 accessions of the TP. The target traits for both TP and CP were DTF, panicle weight (PW) and the nitrogen balance index (NI). Six scenarios, representing different degrees of relatedness between the training set and the progeny set and different sizes of the training set, and three prediction methods, were considered (Table 15.2). In addition, among the six scenarios, three (S1, S2 and S3) were implemented with two different methods of selection of individuals in the training set. Under the first method (a), the training set was composed of accessions of the TP with the lowest average pairwise Euclidian distances with the 31 parental lines of the CP. Under the second method (b), accessions of the training set were

Table 15.2. Scenarios for genomic prediction across generations.

Scenario	Training population	Candidate population
S1a	31 parents	97 progenies
S1b	31 accessions selected using the CDmean method	
S2a	58 related accessions	
S2b	58 accessions selected using the CDmean method	
S3a	31 parents + 58 related accessions	
S3b	89 accessions selected using the CDmean method	
S4	31 parents + 252 accessions	
S5	252 accessions, excluding the parents	
S6	100 random sampling of 31 accessions, excluding the parents	

selected from among the 284 accessions of TP, using the CDmean method of optimization of the training set (Rincent *et al.*, 2012). The selection criterion under CDmean is based on the prediction error variance (PEV) derived from the realized additive relationship matrix–best linear unbiased prediction model. Under the CDmean method, a dedicated training set was selected for each phenotypic trait to account for trait heritability.

The predictive ability of our progeny predictions among the 97 advanced lines of CP was, on average, 0.51 for FL, 0.52 for NI and 0.54 for PW. However, it varied greatly with the composition of the training set (Fig. 15.1). The lower mean PA observed under scenario S1a, compared with scenarios S3a and S4, shows that, in addition to relatedness between the training set and CP, the size of the training set also matters, and even distant accessions can positively contribute to PA of predictions. The results of scenario S2a demonstrate that high PA can be achieved without the presence of the parental lines in the training set, provided it is composed of individuals closely related to the parental lines. The highest PA observed under scenario S4 suggests there is still room for optimization of the size and the composition of the training set; for instance, by selection of the closely related TP individuals, proportional to the contribution of each parental line to the composition of the CP. The almost equal PAs observed in S3a and S4 suggests that, beyond a certain threshold of size of the training set composed of accessions closely related to the CP, the inclusion of less closely related individuals does not improve the PA of prediction. Comparison of PA obtained under scenarios S1a, S2a and S3a, with the accuracy obtained with the corresponding CDmean-based training set (S1b, S2b and S3b) showed almost no gain in predictive ability for DTF and PW, and an almost systematic gain in predictive ability of about 0.1 for NI (data not shown). This is probably attributable to the fact that our scenario for optimization of the training set was also based on relatedness between the training set and the parental lines of the CP. Lastly, but importantly, rather high accuracies (up to 0.7) were obtained among the full-sib lines of individual crosses. However, the number of progenies per cross and the number of crosses analysed were too small to draw general conclusions.

Similar results were obtained in another progeny prediction experiment targeting the improvement of the ability to prevent the accumulation of arsenic (As) in rice grain (Frouin *et al.*, 2019). The problem affects many rice-growing countries and is attributable to the presence of a high concentration of As in the paddy fields (Brammer and Ravenscroft, 2009; Meharg *et al.*, 2009). In this experiment, the TP was composed of 228 *japonica* accessions representing the European Rice Core Collection (Courtois *et al.*, 2012) and the CP was composed of 95 advanced breeding lines developed by the *Centre Français du Riz* (CFR) for the Camargue region, France. The concentration of As in the flag-leaf (FL-As) and in the cargo grain (CG-As) was investigated in a field trial with soil As concentration of 10 mg kg^{-1}. The predictive ability of genomic prediction across populations was evaluated under three scenarios of composition of the training set. Under the first scenario (S1), the training set included all 228 accessions of TP. Under S2, the training set was composed of 100 accessions of the TP, with the lowest average pairwise Euclidian distances with the 95 lines of the CP. Under S3,

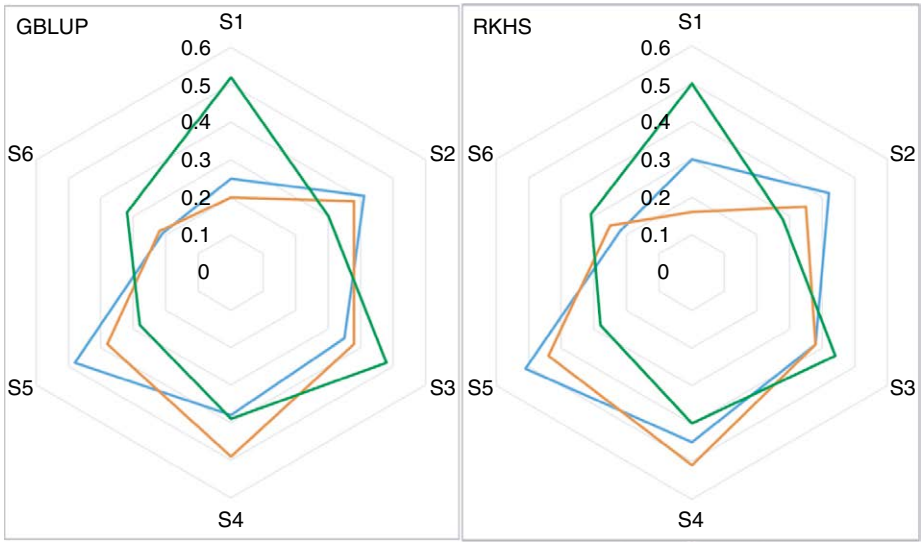


Fig. 15.1. Accuracy of genomic prediction of progeny phenotype for days to flowering (blue), nitrogen balance index (orange) and 100 panicles weight (green), obtained with two statistical methods, genomic best linear unbiased prediction (GBLUP) and reproducing kernel Hilbert space regression (RKHS), under six scenarios (S1–S6) of composition of the training set. See Table 15.2 for details of the six scenarios.

100 accessions of the training set were selected from among the 228 accessions of TP, using the CDmean method.

Under the S1 scenario, the PA of GEBV of the 95 lines of CP was, on average, 0.43 for FL-As and 0.48 for CG-As (Fig. 15.2). The PAs were much lower under S2, compared with S1. Under S3, the average PA was slightly higher than under S1 for CG-As (0.49), and much lower than under S1 for FL-As (0.34). The two prediction methods implemented (GBLUP and RKHS) provided similar levels of PA. However, there was some interaction between prediction methods and phenotypic traits. Translation of the PA observed for CG-As under S3, into average phenotypic performances of CP lines selected on the basis of their GEBV indicates that, for a selection rate of 10%, the difference in genetic gain between phenotypic selection and GEBV-based selection was approximately 5% (Frouin *et al*., 2019).

Integrating Trait-specific Marker Selection

A common feature among almost all published studies on genomic prediction is the use of markers selected on the basis of a variety of criteria except association with the target trait. Prediction models are trained and GEBV are computed using the same set of markers for all the phenotypic traits the breeding programmes is targeting, whatever their genetic architecture. Zhang *et al*. (2010) simulated the predictive ability of different genomic prediction methods trained with a relationship matrix built with markers of equal effect (infinitesimal model) and with the same set of markers with weighted effects. Genomic prediction with markers of weighed effect had higher predictive ability. Similar improvement in predictive ability of genomic prediction for complex traits was reported by Zhang *et al*. (2010), when a trait-specific relationship matrix was built using results of genome-wide association studies (GWAS) available in the literature.

We evaluated the effectiveness of trait-specific marker selection, using a reference population of 204 rainfed lowland accessions with 148,916 SNPs and phenotyped for DTF, PH and GY under three managed environments, E1, E2 and E3. E1 corresponded to the standard lowland rice cultivation, without stress. E2 corresponded to standard lowland rice cultivation associated with application of drought stress at

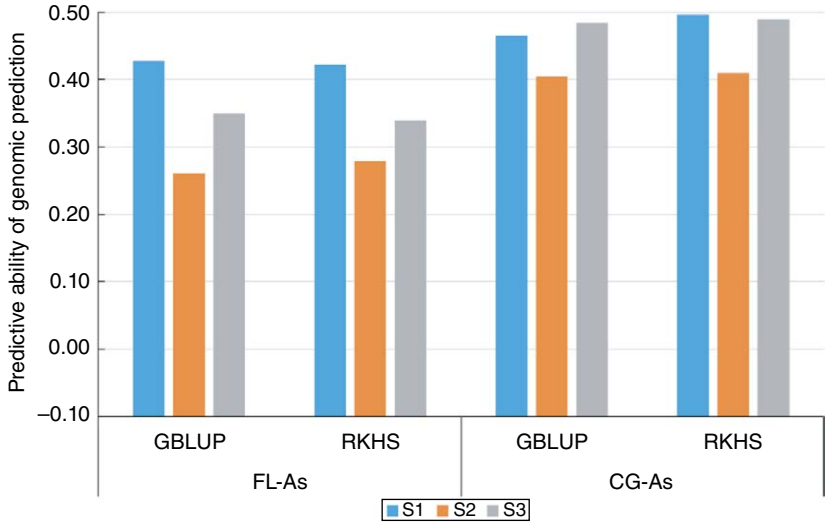


Fig. 15.2. Predictive ability of genomic prediction of the arsenic concentration in the flag leaf (FL-As) and in the cargo grain (CG-As), of the progeny population obtained with two statistical methods, genomic best linear unbiased prediction (GBLUP) and reproducing kernel Hilbert space regression (RKHS), under three scenarios of composition of the training set. S1: TP = 228 accessions; S2: TP = 100 accessions most related with the 95 lines of the CP; S3: TP = 100 accessions selected among the 228, using the CDmean method.

the reproductive stage. E3 corresponded to standard cultivation of upland rice with drought stress applied at the reproductive stage (Bhandari *et al.*, 2019). First, LD-based marker selection was performed with a threshold of $r \leq 0.5$ and MAF ≥ 5%. It yielded 28,091 SNPs. Then, subsets of 28,091 trait-specific SNPs were selected for each trait under each drought environment, using results of GWAS performed with the complete genotypic dataset of 148,916 SNPs. Two scenarios of trait-specific marker selection were considered: (i) implementation of the GWAS experiments with the entire 204 accessions (GWAS-derived markers), and (ii) GWAS experiments with accessions (80% of the total) that did not participate in the corresponding model training process of the genomic prediction experiment (S-GWAS). The prediction models used were GBLUP and RKHS.

Whatever the phenotypic trait, the environment and the genomic prediction method, GWAS-derived markers resulted in systematically significantly higher PA (39% on average) than their LD-derived counterpart (Fig. 15.3). Interactions between the marker selection method, the prediction method and the environment were also often highly significant. For instance, under E1, the average PA gains with GWAS-derived markers, compared with LD-derived markers, were 29% for DTF, 49% for GY and 40% for PH.

The S-GWAS-derived markers also resulted in systematically higher PA of genomic prediction (9% on average) than the corresponding LD-derived markers. The average gain in PA, over the LD-derived markers was 11% under E1, 10% under E2 and 6% under E3. The values of Fisher's least significant difference (LSD) indicate that the gains in PA for DTF, with S-GWAS-derived markers (12% on average), were significant under each of the three environments. For PH, the PA gains (9% on average) were significant only under E1 and E3. For GY, the PA gains (6% on average) were not significant under any of the three environments.

These significant gains in PA suggest that the genetic architecture of the three phenotypic traits considered in our study deviates, to a certain degree, from the infinitesimal model, and each trait is controlled by different sets of QTL. Gains in PA with S-GWAS-based marker selection are less important but more realistic in the context of actual breeding programmes where phenotypic data are not available for individuals

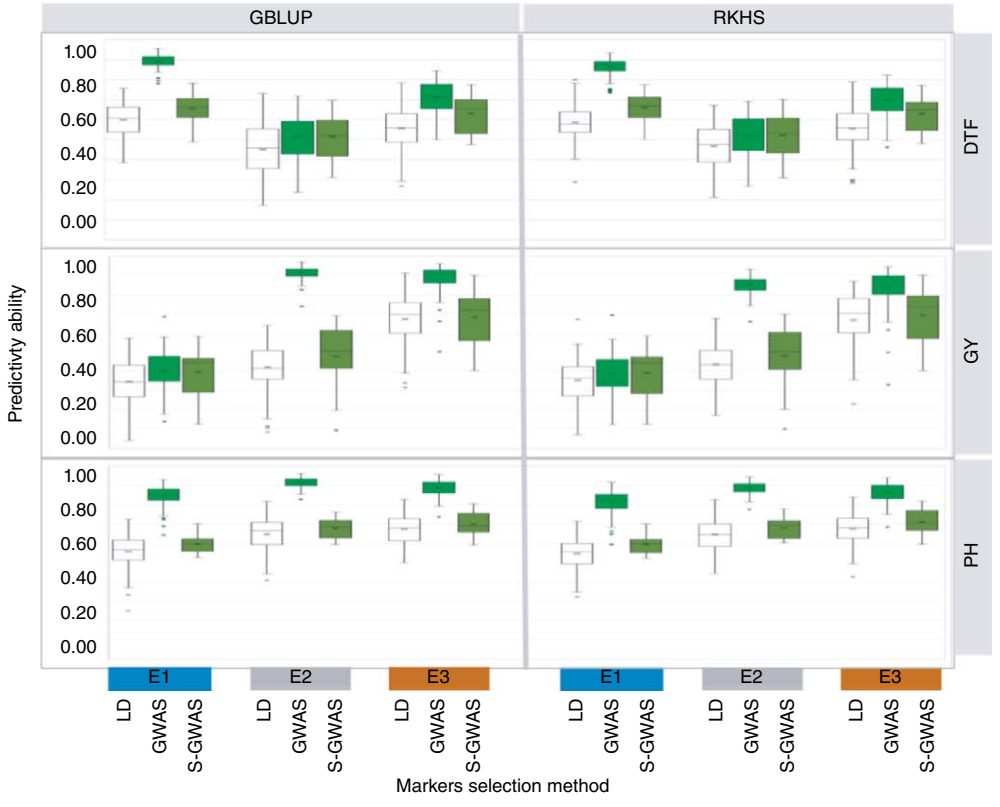


Fig. 15.3. Predictive ability of genomic prediction in cross validation experiments implemented with 28,091 single nucleotide polymorphism (SNP)-derived using three-marker selection methods: linkage disequilibrium (LD) between markers, white boxes; genome-wide association analysis with the target traits (GWAS), pale green boxes; and astringent genome wide association analysis with the target traits (S-GWAS), dark green boxes. The three traits, days to flowering (DTF), grain yield (GY) and plant height (PH), were phenotyped under three environments: rainfed lowland (E1), rainfed lowland with drought stress (E2) and upland with drought stress (E3). For each box, the mean (x) and median (horizontal bar) values are represented.

that are candidates for selection. These gains in PA, despite the limited size of our TP and the rather low heritability of the target traits, which limits the power to detect QTL, suggest that more substantial gains in PA could be achieved if more consolidated QTL information were available. Such consolidated QTL information can be built from the large number of publicly available QTL databases and SNPs detected in different linkage mapping and GWAS experiments. The QTL information could help build trait-specific genomic relationship matrices, based on the modified VanRaden genomic relationship matrix, with marker weights for each locus, as proposed by Zhang *et al*. (2014).

Accounting for GEI to Breed for Tolerance to Abiotic Stresses

In plant breeding, GEI interactions are usually assessed from multi-environment trials and expressed as a change in the relative performance of genotypes in different environments, with or without change in the ranking of the genotypes (Freeman, 1973; Cooper and Hammer, 1996). One specific case of GEI experiments is managed-environment trials that aim to assess the effect of particular environmental variables (e.g. abiotic stresses) or cropping practices (e.g. fertilizer, irrigation, etc.) that influence crop performance in the production environment concerned

(Cooper and Hammer, 1996). A still more specific case of GEI experiments is managed abiotic stress trials that aim to provide a measure of genotypic response to stress based on yield loss under stress compared with normal conditions.

Recently, a number of statistical frameworks that model GEI interactions for the purpose of genomic prediction have been proposed. First, the single-trait-single-environment GBLUP model was extended to a multi-environment context (Burgueño *et al.*, 2012). Then a GBLUP-type model using marker × environment interaction (M×E) was proposed (Lopez-Cruz *et al.*, 2015). Using a non-linear (Gaussian) kernel to model the GEI, the M×E-based approach was further developed (Cuevas *et al.*, 2016). The latest models go beyond the extension of single-environment models and propose multi-environment models based on genetic correlations between environments under two kernel methods, linear (GBLUP) and Gaussian kernel (GK) (Cuevas *et al.*, 2017). Application of these multi-environment models to data from multi-location trials of CYMMYT's maize and wheat breeding programmes confirmed their superiority over the single-environment models. The highest PA of prediction was observed with methods based on genetic correlations between environments.

In rice, we recently reported the first implementation of multi-environment genomic prediction, in the context of managed abiotic stress trials (Ben Hassen *et al.*, 2018). The above-described TP of 284 accessions was phenotyped not only under the conventional continuous flooding but also under the more water-sparing system of alternate wetting and drying. The PAs of multi-environment models were compared with the PA of single-environment models for DTF, NI and PW, under two cross validation strategies: CV1 where it is assumed that no phenotypic data are available for the CP accessions; and CV2, where it is assumed that phenotypic data for the TP accessions are available under one of the two environments. The extended GBLUP model (Lopez-Cruz *et al.*, 2015) and the extended RKHS model (Cuevas *et al.*, 2017) that integrate environmental effects were used to predict the GEBV with data from the two water-management systems. In the extended GBLUP model, the effects of m environments and the effects of P markers are separated into two components: the main effect of the markers for all the environments and the effect of the markers for each environment. In the extended RKHS model, the mixed model is written as follows:

$$y = \mu + u + f + \varepsilon$$

where y is the response vector, μ is the vector with the intercept of each environment, u the random vector of individual genetic values, f the genetic effects associated with individuals that were not accounted for in component u, and ε the random vector of the error. u, f and ε are independent and normally distributed.

Average predictive abilities ranged from 0.48 to 0.96, depending, in decreasing importance, on the trait, the type of model (i.e. single- versus multi-environment), the cross validation strategy, the statistical model and the water-management system (Fig. 15.4). The average predictive ability was 0.77, 0.56 and 0.68 for FL, NI and PW, respectively. Whatever the trait or the water-management system, multi-environment models with the CV1 strategy performed similarly to the single-environment model. Conversely, the multi-environment models with the CV2 strategy outperformed single-environment models with an average gain of 0.27 for FL, 0.12 for NI and 0.20 for PW. Among the multi-environment prediction models, RKHS-2 performed systematically slightly better than GBLUP, with a gain in predictive ability of up to 0.02. We confirmed the superiority of multi-environment models over single-environment models, in the context of breeding for drought tolerance for the rainfed, lowland rice-growing environments (Bhandari *et al.*, 2019).

Implications for Breeding Rice

It is now widely accepted that the prediction ability for complex traits is better when using whole-genome marker prediction than when using a few markers targeting a few QTL (Crossa *et al.*, 2017; Hickey *et al.*, 2017). The general rules for the implementation of GS-based plant breeding are also well established, thanks to extensive simulation and empirical data-based analyses of the effects of factors that affect the

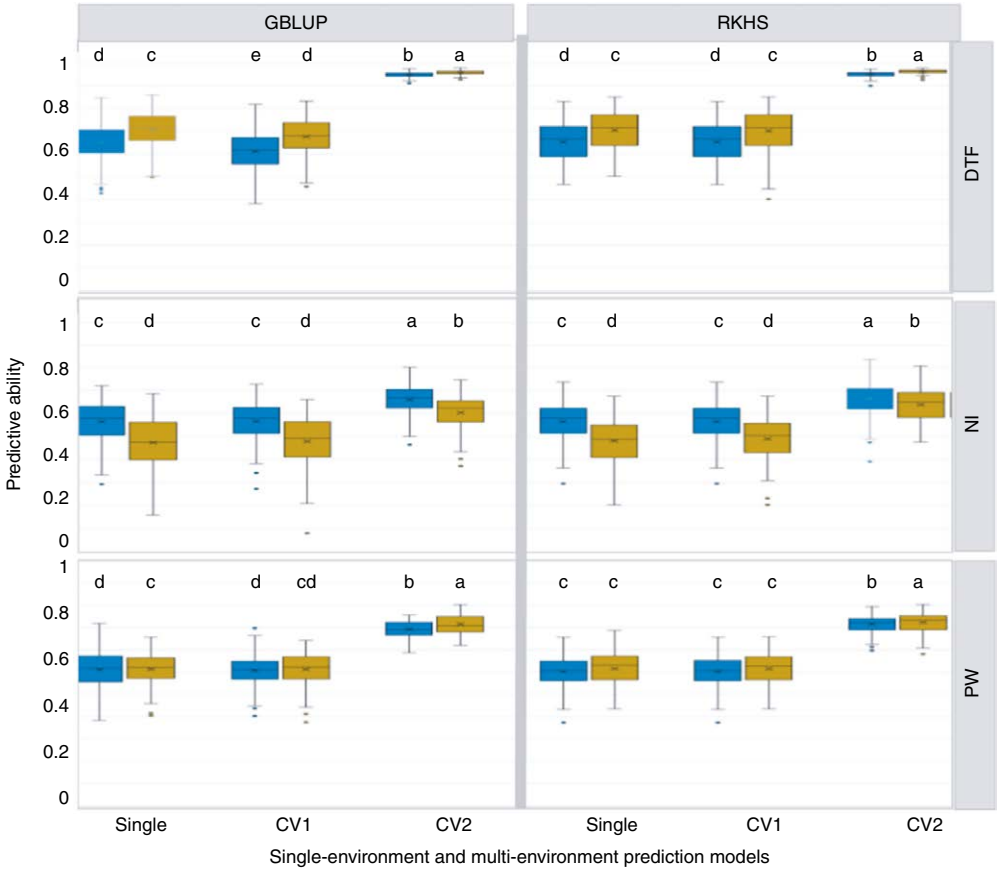
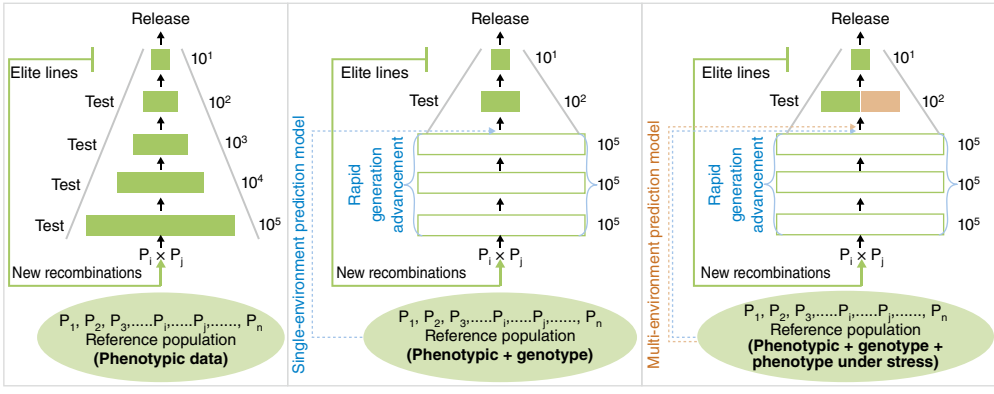


Fig. 15.4. Single-environment and multi-environment (CV1 and CV2) predictive ability in cross validation experiments in the reference population obtained with three statistical models (genomic best linear unbiased prediction [GBLUP], reproducing kernel Hilbert space regression [RKHS]). Continuous flooding and alternate wetting and drying water management conditions are in blue and orange, respectively. The three studied traits presented are: days to flowering (DTF), nitrogen balance index (NI) and panicle weight (PW). The letters in each panel represent the results of Tukey's HSD comparison of means and apply to each panel independently. The means differ significantly (*P* value < 0·05) if two boxplots have no letter in common.

PA of GEBV in all major crops, including rice. We recently undertook a number of studies using data from ongoing rice breeding programmes to draw decision-making rules for the most popular rice breeding scheme and for the rather common breeding objective of yield potential and tolerance to abiotic stresses.

Pedigree breeding within the progenies of biparental crosses extracted from a working collection or reference population (Fig. 15.5a) is the most common scheme for the improvement of complex traits in rice, as in many other autogamous crops (Bernardo, 2014). We found that using phenotypic and genotypic data from the reference population to train the prediction model made it possible to predict performances among the first generation of advanced (F_5–F_7) progeny of a large set of biparental crosses. Thus, breeders can use this prediction approach in the framework of a pedigree breeding scheme. The approach can be associated with rapid generation advancement (in off-season nurseries or controlled environments), a practice aimed at reducing the length of the breeding cycle and hence accelerating the genetic gain per unit of time (Fig. 15.5b; Collard *et al.*, 2017). Specific optimization of the training set might be needed to obtain the best possible



(a) Conventional pedigree breeding

(b) Genomic pedigree breeding for favourable environment

(c) Genomic pedigree breeding for favourable and stress environments

Fig. 15.5. Schematic representation of implementation of genomic selection in the framework of pedigree breeding. Shade of green colour of the reference population ellipses represents the increasing amount of phenotypic and genotypic data needed.

prediction accuracy for the progeny of each cross. The scheme can also be applied when homozygous lines from biparental crosses are produced by haplodiploidization, at least in the *japonica* genetic group, for which a high-throughput method is available (Alemanno and Guiderdoni, 1994). As the GS-based advanced line will then go through two or three cycles of phenotypic evaluation, the data collected will provide an opportunity to further refine the training model (Heffner *et al.*, 2010).

The high predictive ability of multi-environment, genomic prediction we observed, in two managed abiotic-stress case studies, paves the way for a new breeding option: conducting simultaneously direct and indirect selection for performance under both stress and non-stress environments. It requires a training population carefully phenotyped under both favourable environment and managed drought. While, in a first step of selection, the candidate population would be phenotyped only under the less expensive favourable environments (Fig. 15.5c), the selected candidate would be phenotyped under the target stress environment to ascertain their GEBV and to update the multi-environment prediction model for the next breeding cycle (Heffner *et al.*, 2009; Pszczola and Calus, 2016). The process can be implemented in the framework of the pedigree-breeding of progeny derived from biparental crosses between members of the reference population of the breeding programme, which is used as a training population, as shown by Ben Hassen *et al.* (2018).

The effectiveness of trait-specific markers, in the context of multi-environment, model-based genomic prediction, deserves investigation using the simulation approach. Nevertheless, breeders should consider the inclusion of a limited share of trait-specific markers (especially for the most important target traits) when genotyping candidate populations.

The next step in harnessing the potential of GS in rice breeding would be the wider adoption of a population improvement scheme (Guimaraes, 2005) that allows gradual increase in frequency of favourable alleles and ensures better maintenance of QTL-marker LD along the breeding cycles, and thus the persistence of the prediction model. Such schemes associated with GS models that predict the line value of heterozygous individuals (Gallais, 1979) as early as S_0 generation would accelerate genetic gain by reducing the length of breeding cycles and providing an opportunity for increasing selection pressure. Our team has undertaken the development of such a model in the framework of the upland rice breeding programme we are running in collaboration with CIAT in Colombia, using a population improvement scheme (Grenier *et al.*, 2015). Another important step in harnessing the potential of GS in rice breeding would be connecting the genomic prediction models with ecophysiological crop models, such

as EcoMeristem (Luquet *et al*., 2006), and Samara (Dingkuhn *et al*., 2016) to predict GEI for unobserved environments, and thus the performances and stability of the lines, based only on the genotypic information.

Summary

The increase in rice production needed to meet future demand requires new cropping systems and rice varieties with enhanced resource-use efficiency and adaptation to environmental stresses in the context of climate change. GS has the potential to accelerate the development of such varieties. We present a review of literature analysing factors that affect the predictive ability of genomic prediction in rice, and an overview of the proof-of-concept studies conducted during the past 5 years by our team with the aim of providing rice breeders with tailored GS methods and tools. These studies involved two complementary breeding schemes (pedigree breeding and population improvement), mobilized different compartments of the rice genetic diversity (*indica*, tropical *japonica* and temperate *japonica*) and targeted a wide range of traits (yield potential, adaptation to alternate watering and drying, drought tolerance and exclusion of heavy metals). Issues addressed include training the population for making selection decisions in pedigree breeding within the progeny of biparental crosses, accounting for information available on gene/QTL involved in the determination of complex traits and accounting for genotype-by-environment interactions. In the light of the results of these studies, we discuss a strategy for the implementation of genomic selection in the framework of pedigree breeding. We conclude on issues that need further simulation and empirical studies to fully harness the potential of GS.

References

Alemanno, L. and Guiderdoni, E. (1994) Increased doubled haploid plant regeneration from rice (*Oryza sativa* L.) anthers cultured on colchicine-supplemented media. *Plant Cell Reports* 13, 432–436.

Ben Hassen, M., Cao, T.-V., Bartholomé, J., Orasen, G., Colombi, C., *et al*. (2017) Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. *Theoretical and Applied Genetics* 131(2), 417–435. DOI: 10.1007/s00122-017-3011-4.

Ben Hassen, M., Bartholomé, J., Valè, G., Cao, T.-V. and Ahmadi, N. (2018) Genomic prediction accounting for genotype by environment interaction offers an effective framework for breeding simultaneously for adaptation to an abiotic stress and performance under normal cropping conditions in rice. *Genes, Genomes, Genomics* 8(9), 2319–2332. DOI: 10.1534/g3.118.200098.

Bernardo, R. (2014) *Essentials of Plant Breeding*. Stemma Press, Woodbury, Minnesota.

Bhandari, A., Bartholomé, J., Cao, T.-V., Kumari, N., Frouin, J., *et al*. (2019) Selection of trait-specific markers and multi-environment models improve genomic predictive ability in rice. *PLoS ONE* 14(5), e0208871. DOI: 10.1371/JOURNAL.PONE.0208871.

Brammer, H. and Ravenscroft, P. (2009) Arsenic in groundwater: A threat to sustainable agriculture in South and South-east Asia. *Environment International* 35, 647–654.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. (2012) Genomic prediction of breeding values when modelling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science* 52(2), 707–719.

Collard, B., Beredo, J., Lenaerts, B., Mendoza, R., Santelices, R., *et al*. (2017) Revisiting rice breeding methods – evaluating the use of rapid generation advance (RGA) for routine rice breeding. *Plant Production Science* 20(4) 337–352. DOI: 10.1080/1343943X.2017.1391705.

Cooper, M. and Hammer, G.L. (1996) *Plant Adaptation and Crop Improvement*. CAB International, Wallingford, UK.

Courtois, B., Frouin, J., Greco, R., Bruschi, G., Droc, G., *et al*. (2012) Genetic diversity and population structure in a European collection of rice. *Crop Science* 52, 1663–1675.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., *et al*. (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* 22(11), 961–975.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., *et al*. (2016) Genomic prediction of genotype × environment interaction kernel regression models. *The Plant Genome* 9(3), 1–20. DOI: 10.3835/plantgenome2016.03.0024.

Cuevas, J., Crossa, J., Montesinos-López, O.A., Burgueño, J., Pérez-Rodríguez, P., *et al*. (2017) Bayesian genomic prediction with genotype × environment interaction kernel models. *Genes, Genomes, Genetics* 7(1), 41–53.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. and Calus, M.P.L. (2015) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 93, 327–345.

Dingkuhn, M., Kumar, U., Laza, M., Pasco, R. (2016) SAMARA: A crop model for simulating rice phenotypic plasticity. In: Ewert, F., Boote, K.J., Rötter, R.P., Thorburn, P. and Nendel, C. (eds) *Crop Modelling for Agriculture and Food Security Under Global Change. Abstracts*. International Crop Modelling Symposium, Berlin, Germany, 15–17 March 2016, pp. 45–46.

Freeman, G.H. (1973) Statistical methods for the analysis of genotype-environment interactions. *Heredity* 31(3), 339–354.

Frouin, J., Labeyrie, A., Boisnard, A., Sacchi, G.A. and Ahmadi, N. (2019) Genomic prediction offers the most effective marker assisted breeding approach for ability to prevent arsenic accumulation in rice grains. *PLoS ONE* 14(6), e0217516. DOI: 10.1371/journal.pone.0217516.

Gallais, A. (1979) The concept of varietal ability in plant breeding. *Euphytica* 28(3), 811–823.

Gezan, S.A., Osorio, L.F., Verma, S. and Whitaker, V.M. (2017) An experimental validation of genomic selection in octoploid strawberry. *Horticulture Research* 4, 16070. DOI: 10.1038/hortres.2016.70.

Goddard, M.E. and Hayes, B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* 10, 381–391.

Grenier, C., Cao, T.-V., Ospina, Y., Quintero, C., Châtel, M.H., *et al*. (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PLoS ONE* 10(8), e0136594. DOI: 10.1371/journal.pone.0136594.

Guimaraes, E.P. (2005) *Population Improvement: A Way of Exploiting the Rice Genetic Resources of Latin America*. FAO, Rome, pp. 56–94.

Guo, Z., Tucker, D.M., Basten, C.J., Gandhi, H., Ersoz, E., *et al*. (2014) The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics* 127, 749–762. DOI: 10.1007/s00122-013-2255-x.

Hayes, B.J., Bowman, P., Chamberlain, A., Verbyla, K. and Goddard, M. (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* 41, 51.

Heffner, E.L., Sorrells, M.E. and Jannink, J.-L. (2009) Genomic selection for crop improvement. *Crop Science* 49(1), 1–12.

Heffner, E.L., Lorenz, A.J., Jannink, J.-L. and Sorrells, M.E. (2010) Plant breeding with genomic selection: Gain per unit time and cost. *Crop Science* 50, 1681–1690. DOI: 10.2135/cropsci2009.11.0662.

Heslot, N., Sorrells, M.E., Jannink, J.-L. and Yang, H.P. (2012) Genomic selection in plant breeding: A comparison of models. *Crop Science* 52, 146–160.

Hickey, J.M., Chiurugwi, T., Mackay, I. and Powell, W. (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics* 49(9), 1297–1303.

Hofheinz, N., Borchardt, D., Weissleder, K. and Frisch, M. (2012) Genome based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and Applied Genetics* 125, 1639–1645. DOI: 10.1007/s00122-012-1940-5.

Howard, H., Carriquiry, A.L. and Beavis, W.D. (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *Genes, Genomes, Genomics* 4, 1027–1046.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., *et al*. (2015) Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics* 128, 145–158. DOI: 10.1007/s00122-014-2411-y.

Jannink, J.L., Lorenz, A.J. and Iwata, H. (2010) Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics and Proteomics* 9, 166–177. DOI: 10.1093/bfgp/elq001.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., *et al*. (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *Genes, Genomes, Genetics* 5(4), 569–582.

Lorenzana, R.E. and Bernardo, R. (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120, 151–161.

Luquet, D., Dingkuhn, M., Kim Hae, K., Tambour, L. and Clément-Vidal, A. (2006) EcoMeristem, a model of morphogenesis and competition among sinks in rice: 1. Concept, validation and sensitivity analysis. *Functional Plant Biology* 33(4), 309–323. DOI: 10.1071/FP05266.

Meharg, A.A., Williams, P.N., Domako, E.A., Lawgali, Y., Deacon, C., *et al*. (2009) geographical variation in total and inoganic arsenic content of polished (white) rice. *Environmental Science & Technology* 43, 1612–1617.

Meuwissen, T.H.E. (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics Selection Evolution* 41, 35.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4), 1819–1829.

Onogi, A., Ideta, O., Inoshita, Y., Ebana, K., Yoshioka, T., *et al*. (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* 128, 41–53. DOI: 10.1007/s00122-014-2411-y.

Pszczola, M. and Calus, M.P. (2016) Updating the reference population to achieve constant genomic prediction reliability across generations. *Animal* 10(6), 1018–1024. DOI: 10.1017/S1751731115002785.

Riedelsheimer, C., Endelman, J.F., Stange, M., Sorrells, M.E., Jannink, J.-L., *et al*. (2013) Genomic predictability of interconnected biparental maize populations. *Genetics* 194, 493–503.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., *et al*. (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192(2), 715–728. DOI: 10.1534/genetics.112.141473.

Sallam, A., Endelman, J., Jannink, J.L. and Smith, K. (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome* 8, 1. DOI: 10.3835/plantgenome2014.05.0020.

Spindel, J., Begum, H., Virk, P., Collard, B., Redoña, E., *et al*. (2015) Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics* 11, e1004982. DOI: 10.1371/journal.pgen.1004982.

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., *et al*. (2017) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity* 118, 302–310.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., *et al*. (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5(9), e12648. DOI: 10.1371/journal.pone.0012648.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., *et al*. (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* 9(3), e93017. DOI: 10.1371/journal.pone.0093017.

16 Novel Breeding Approaches for Developing Climate-resilient Rice

Sandeep Chapagain, Lovepreet Singh and Prasanta K. Subudhi*
Louisiana State University Agricultural Center, Baton Rouge, Louisiana, USA

Introduction

Abiotic stresses are major impediments to rice productivity in almost all production environments. Since more farmland is affected by various abiotic constraints resulting from climatic disturbances, achieving food security has been an arduous task. As a staple food for the majority on this planet, it is essential to enhance rice production for feeding the exponentially growing world population. Under adverse environmental conditions, crops exploit only part of their genetic potential (Boyer, 1982). Different environmental constraints such as drought, flooding, salinity and extreme temperature profoundly affect plant growth and productivity. Crops require a variety of mechanisms to adapt and survive under an unfavourable environment. The development of cultivars with increased abiotic stress tolerance is a logical approach to improving rice production. Although abundant genetic variation exists in the rice gene pool, progress is slow because of the complex genetics of the tolerance mechanisms and low heritability associated with abiotic stress tolerance. The development of new genomic tools offers unique promise to design climate-resilient rice through enhanced understanding of various tolerance mechanisms. Both conventional and molecular tools along with knowledge emanating from various 'omics' approaches can be used to exploit the natural variability present in the world germplasm for developing rice cultivars with enhanced yield under a range of environmental constraints.

Abiotic Stresses and Tolerance Mechanisms in Rice

Rice is a salt-sensitive crop with a threshold of 3 dS m^{-1} (Maas, 1990). It is highly vulnerable to salinity stress at both seedling and reproductive stages. Plants experience both osmotic and ionic stresses in saline environments. Salinity reduces the rate of net CO_2 assimilation, transpiration rate, stomatal conductance, relative water content, water use efficiency (WUE), dry matter accumulation, leaf development and pollen fertility (Hussain *et al.*, 2017).

There are three mechanisms that rice plants exploit to tolerate salinity stress: osmotic tolerance, tissue tolerance and ion exclusion (Munns and Tester, 2008). In osmotic tolerance, plants maintain stomatal conductance and leaf expansion. Tissue tolerance includes synthesis of compatible solutes, sequestration of Na^+ in the vacuole and enzyme production for

* Email: psubudhi@agcenter.lsu.edu

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

detoxification of reactive oxygen species (ROS). Ion exclusion involves the transport of Na^+ from xylem to soil. Tolerant lines have low sodium ion accumulation, and increased absorption of K^+, resulting in low Na^+/K^+ ratios. Ion pumps present in cell membranes like antiporters, symporters and carrier proteins maintain cell ion homeostasis (Blumwald, 2000).

Drought stress causes various morphological, biochemical and physiological changes in rice plants which ultimately reduce yield (Ji *et al*., 2012). Typically, drought stress during grain filling brings early senescence and reduces grain-filling time. The physiological impacts include reduction in photosynthesis, transpiration rate, stomatal conductance, WUE, relative water content and membrane stability index. Rice plants also accumulate various organic and inorganic compounds in the cytosol to maintain cell turgidity under drought stress.

Plants adapt to drought stress by three main mechanisms: drought tolerance, drought avoidance and drought escape (Fukao and Xiong, 2013). Drought escape allows rice plants to produce more grains despite limited water supply. The plant's ability to maintain high tissue water potential under a moisture stress condition is termed as drought avoidance. Rice genotypes having deep, coarse roots with higher branching ability and soil penetration, high root–shoot ratio, reduced leaf rolling, early stomata closure and higher cuticular resistance can avoid drought. Drought tolerance allows the plant to survive or grow under limited water content in the tissue.

Flooding is a major constraint in rainfed lowland and irrigated rice-growing regions of South and South-east Asia. Although rice survives in a waterlogged environment because of interconnected aerenchyma tissues facilitating aeration in root tissues, most rice varieties do not tolerate flooding for more than a week (Singh *et al*., 2017). There are many root and shoot traits that enable rice plants to survive under flooding and some important traits include transient dormancy of shoots, shoot elongation ability and underwater photosynthesis (Ismail, 2018).

In a current climatic scenario, an increase in temperature above the optimal range is frequent. Rice is highly susceptible to high temperature (Rabara *et al*., 2018). The optimum temperature for rice is 22–30°C. High temperature results in reduced grain weight, grain filling, grain size, increase in chalkiness and reduction in amylose content (Jagadish *et al*., 2010). An increase in temperature of 2°C results in drastic reduction in rice yield, and genotypes which flower early in the morning and/or have large anthers are more tolerant to heat (Shah *et al*., 2014). Some physiological heat tolerance mechanisms include a change in photosynthetic rate, higher accumulation of heat-shock proteins (HSPs), oxidants and osmoprotectants, change in leaf position, reduction in transpiration, and changes in hormone levels and metabolites.

Rice is susceptible to damage by temperatures below 15°C (Howarth and Ougham, 1993). Low-temperature stress causes significant damage at the booting stage (Pan *et al*., 2015). Chilling stress causes poor germination, seedling injury, poor crop establishment and reduction in yield (Andaya and Mackill, 2003). Physiological changes include increased electrolyte leakage, lipid peroxidation, proline and other metabolites, and changes in chlorophyll fluorescence. The rigidity of the plasma membrane has been shown to induce cold-responsive genes to enhance cold tolerance during cold stress (Örvar *et al*., 2000). The unsaturated fatty acids in the plasma membrane prevent electrolyte leakage and cell death, resulting in cold tolerance. Better understanding of various tolerance mechanisms through application of genetics and new omic technologies will help accelerate breeding rice varieties with improved abiotic stress tolerance.

Conventional and Mutation Breeding

In conventional plant breeding, various selection strategies have been adopted to develop abiotic stress-tolerant rice varieties. Salinity-tolerant varieties CSR-1, CSR-2 and CSR-3 were developed by pure line selection from locally grown cultivars in India. Similarly, other varieties (SR-26B, Hamilton, Patnai-23 and Jhona-349) were developed by site-specific selection in different countries in improved backgrounds (Gregorio *et al*., 2002). Backcross breeding was used to develop introgression lines with salt tolerance (Puram *et al*., 2018) and tolerance to multiple abiotic stresses (Ali *et al*., 2017). De Leon *et al*. (2015) evaluated US rice varieties and several

imported germplasms for salinity tolerance in seedling stage and suggested Geumgangbyeo and TCCP266 as potential donors for use in breeding programmes.

Mutagenesis has been used to create new desirable genetic variability as well as to improve well-adapted cultivars only deficient in one or two traits such as salt and drought susceptibility (Hallajian, 2016). Both salt-tolerant and drought-tolerant rice cultivars have been developed using this approach (Saleem *et al.*, 2005; Hay *et al.*, 2015).

Novel Approaches to Improve Abiotic Stress Tolerance in Rice

Because of the genetic complexity of abiotic stress tolerance in plants, novel approaches have been developed and employed to discover the genes involved in adaptation in abiotic stress environments as well as for development of climate-resilient rice varieties (Fig. 16.1). These approaches involve various omics tools, genomic selection, marker-assisted selection and genome engineering using transgenesis and genome editing. 'Omics' technologies, which deal with genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolite (metabolomics), are being employed for enhanced understanding of abiotic stress tolerance.

Genomics

Genomics offers unparalleled possibilities to understand the molecular basis of abiotic stress tolerance. It encompasses tools such as mapping and sequencing to dissect the complex abiotic stress-tolerance traits, leading to cloning of genetic determinants and marker-assisted selection for developing climate-resilient varieties.

QTL mapping

Quantitative trait loci (QTL) analysis integrates phenotypic and genotypic data to understand the genetic basis of variation in quantitative traits. The number and location of QTL for traits associated with abiotic stress tolerance, followed by gene discovery and marker-assisted selection, are accelerating genetic progress in the development of abiotic stress-tolerant plants.

SALINITY Several QTL-mapping studies have been conducted in populations involving different donors (reviewed by Karan and Subudhi, 2011; Negrão *et al.*, 2011). Among these QTL, the seedling stage tolerance QTL, *Saltol*, is useful because of its large effect (Bonilla *et al.*, 2002). Another QTL for salt tolerance, *SKC1*, which maintains K^+ homeostasis, was cloned (Ren *et al.*, 2005). De Leon *et al.* (2016) identified 16 large effect QTL for nine traits for seedling stage salinity tolerance using a high-density SNP linkage map generated by genotyping-by-sequencing (GBS) in a recombinant inbred line (RIL) population from a cross involving a salt-tolerant donor 'Pokkali'. Gimhani *et al.* (2016) mapped 83 QTL for 11 morphophysiological traits associated with seedling stage salinity tolerance in an RIL population derived from the cross At354 × Bg352 and the majority were clustered in 14 genomic regions. Introgression line populations of the salt-tolerant donors Pokkali and Nona Bokra, developed in several US cultivar backgrounds,

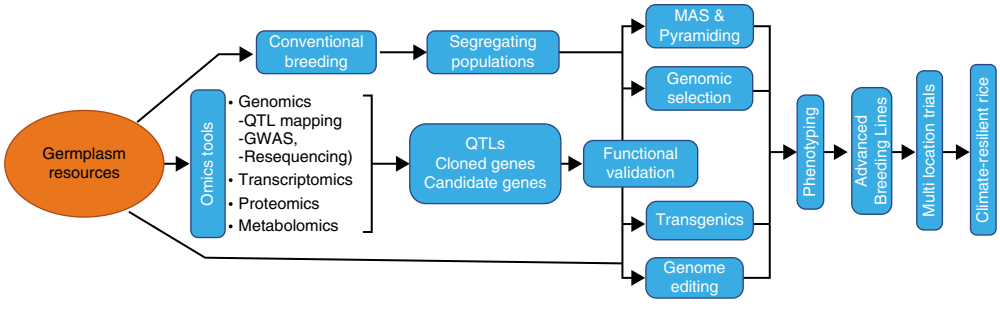


Fig. 16.1. A scheme for integrating novel breeding tools to develop climate-resilient rice varieties.

not only identified new QTL for seedling stage tolerance but also validated many of the earlier identified QTL (De Leon *et al*., 2017; Puram *et al*., 2017, 2018). Mohammadi *et al*. (2013) detected 35 QTL for salinity tolerance at reproductive stage in a cross between Sadri and FL478, and major QTL were located on chromosomes 2, 4 and 6. Bulk segregant analysis of RILs was used to identify three salt-tolerant QTL at reproductive stage for grain yield (Tiwari *et al*., 2016). Bizimana *et al*. (2017) identified 20 new QTL for different traits on chromosomes 1, 2, 4, 6, 8 and 9 using an RIL population of the cross Hasawi × IR29 but Hasawi did not contain the same salinity-tolerance allele as Nona Bokra and Pokkali at *SKC1* and *Saltol*. Furthermore, the same population was evaluated in different environments leading to identification of common regions for multiple QTL on chromosomes 1, 4, 6, 8 and 12, suggesting the presence of novel salt-tolerance alleles in Hasawi (Rahman *et al*., 2017). The meta-analysis of QTL related to seedling salinity tolerance revealed many candidate genes that can be helpful for marker-assisted selection, fine mapping, pyramiding and cloning to understand the salinity-tolerance mechanism in rice (Islam *et al*., 2019). Takagi *et al*. (2015) discovered a gene for salt tolerance using MutMap (Abe *et al*., 2012) and developed a salt-tolerant variety named 'Kaijin'.

DROUGHT Earlier QTL studies largely focused on root traits such as length, thickness, penetration ability, root-pulling force, seminal root length, adventitious root number, and length and number of lateral roots (Zhang *et al*., 2001a,b). Bhattarai and Subudhi (2018a) mapped 14 additive-effect QTL for root and shoot traits in the cross Cocodrie × N22, and the majority of these were clustered in chromosome 1. A QTL for root growth angle *DRO1* was cloned to improve drought avoidance (Uga *et al*., 2013).

Recent studies have largely focused on mapping QTL for yield and yield component traits under drought stress. Several major grain yield QTL were mapped under drought conditions (reviewed by Kumar *et al*., 2014, 2018). The identified QTL were consistent in multiple genetic backgrounds and various target environments (Vikram *et al*., 2011; Prince *et al*., 2015). Twenty-one QTL were discovered for heading date, plant height, leaf rolling score, plant dry matter content, spikelet fertility, grain yield, yield index and harvest index under drought stress at the reproductive stage in an RIL population from the cross Cocodrie × N22 (Bhattarai and Subudhi, 2018b). The drought-yield QTL, *qDTY3.2* from Moroberekan × Swarna was associated with early flowering, reduction of shallow root length and deep root growth (Grondin *et al*., 2018). Catolos *et al*. (2017) mapped three major yield QTL and two QTL for a root development trait in a mapping population derived from a cross between Dular and IR64-21.

SUBMERGENCE The genetic basis of submergence tolerance has been investigated by many researchers (reviewed by Singh *et al*., 2017). Most notable among them is the cloning of a major QTL *Sub1* from the tolerant variety FR13A and it contained three ethylene response factor (ERF) genes: *Sub1A, Sub1B* and *Sub1C* (Xu *et al*., 2006). Hattori *et al*. (2009) discovered two ERF genes, *SNORKEL1* (*SK1*) and *SNORKEL2* (*SK2*), in the QTL region for the internode elongation ability. In addition, *Leaf Gas Film 1* (*LGF1*) gene, with a role in gas exchange and underwater photosynthesis, was isolated (Kurokawa *et al*., 2018). Since tolerance to anaerobic germination is an important attribute for direct seeded rice, trehalose-6-phosphate phosphatase (*OsTPP7*) was identified as the causal factor underlying the major QTL *qAG9-2* for this trait (Kretzschmar *et al*., 2015).

HEAT STRESS Several studies have focused on genetic dissection of heat tolerance during the reproductive stage because of its negative impact on grain quality (Ye *et al*., 2015; Shanmugavadivel *et al*., 2017; Zhu *et al*., 2017). Two major QTL, *qHTSF1.1* and *qHTSF4.1*, were mapped in a IR64/N22 population, and subsequent fine mapping of *qHTSF4.1* that improved spikelet fertility by about 15% under heat stress suggested that cell wall-associated kinase genes may be responsible for heat tolerance. An RIL population from the cross N22/IR64 led to discovery of multiple QTL for spikelet sterility and yield under heat stress using SNP markers (Shanmugavadivel *et al*., 2017). Liu *et al*. (2017) mapped four major QTL for heat stress tolerance at the flowering stage in an introgression line population of a heat-tolerant variety Gan-Xiang-Nuo and identified eight genes in the QTL regions by integrating small RNA sequencing with QTL mapping information. Kilasi *et al*. (2018) focused on vegetative stage heat stress tolerance of 'N22'

and identified six QTL for root length and two for shoot length.

COLD STRESS Genetics of cold tolerance at both seedling and reproductive stages has been investigated in several QTL mapping studies (reviewed by Zhang *et al.*, 2014) and several cold tolerance loci were genetically characterized (Ma *et al.*, 2015; Zhang *et al.*, 2017; Zhao *et al.*, 2017; Liu *et al.*, 2018; Xiao *et al.*, 2018). Recently, Mao *et al.* (2019) cloned a QTL, *HAN1*, conferring cold tolerance in temperate *japonica* rice and demonstrated that an insertion of a MYB cis element in the promoter of this gene was responsible for adaptation in temperate rice growing areas. Bulked segregant analysis (BSA) was coupled with next-generation sequencing strategy to identify cold tolerance QTL and the candidate genes (Sun *et al.*, 2018).

Genome-wide association studies

Genome-wide association studies (GWAS) have been successfully applied for genetic analysis of tolerance to drought (Ma *et al.*, 2016; Pantalião *et al.*, 2016), salinity (Ahmadi *et al.*, 2011; Kumar *et al.*, 2015), chilling tolerance (Pan *et al.*, 2015; Pandit *et al.*, 2017; Schläppi *et al.* 2017) and submergence tolerance (Zhang *et al.*, 2017). Lekklar *et al.* (2019) identified 146 genes which are co-localized with earlier reported QTL for salinity tolerance at the flowering stage in a panel of Thai varieties. A GWAS study on drought tolerance during the vegetative stage of rice identified 39 QTL for different traits in Vietnamese landraces (Hoang *et al.*, 2019). Al-Tamimi *et al.* (2016) evaluated transpiration rate, relative growth and transpiration use efficiency (TUE) from *indica* and *aus* rice panels under salinity and waterlogged conditions and identified previously undetected loci for TUE on chromosome 11. Swamy *et al.* (2017) detected 7 marker trait associations (MTA) for grain yield under drought stress. It is a useful tool for discovering new genes and alleles for complex traits like abiotic stress tolerance from diverse germplasm.

Marker-assisted selection

Marker-assisted selection (MAS) is a strategy to accelerate genetic gain in conventional breeding programmes by selecting plants with a desirable combination of genes using tightly linked markers. It can be exploited to introgress large-effect QTL or pyramid multiple QTL for developing improved abiotic stress-tolerant rice varieties (Kumar *et al.*, 2018). Several markers have been developed for MAS in breeding programmes to enhance tolerance to abiotic stresses (Nogoy *et al.*, 2016; Das *et al.*, 2017).

Marker-assisted backcross breeding is becoming a popular strategy to enhance grain yield under different abiotic stresses. Several studies have reported that QTL pyramiding increased grain yield and tiller formation under drought stress conditions (Kumar *et al.*, 2014; Shamsudin *et al.*, 2016; Anyaoha *et al.*, 2019). Marker-assisted backcrossing has been used to incorporate the large effect salt tolerant QTL '*Saltol*' into popular varieties, which showed enhanced tolerance to salinity at the seedling stage (Punyawaew *et al.*, 2016; Singh *et al.*, 2016; Valarmathi *et al.*, 2019).

Many popular varieties were improved for submergence tolerance by transferring *Sub1* using MAS (Dar *et al.*, 2018). Das and Rao (2015) stacked *Saltol* and *Sub1* along with several biotic stress-tolerance genes into a rice variety. Marker-assisted pyramiding of multiple drought tolerance QTL and *Sub1* led to the development of drought- and submergence-tolerant varieties without any yield penalty in non-stress environments (Sandhu *et al.*, 2019). Dharmappa *et al.* (2019) used marker-assisted backcross breeding to transfer root and water use efficiency traits to develop improved IR64, which performed well under semi-irrigated conditions. Cold-tolerant lines were developed by a QTL pyramiding approach in 'Hitomebore' and '93-11' backgrounds (Endo *et al.*, 2016; Li and Mao, 2018).

Genomic selection

Genomic selection (GS) is an approach used to improve the efficiency of breeding effort involving quantitative traits. It involves the prediction of the breeding value of each individual compared to the identification of QTL for use in a traditional MAS programme (Asoro *et al.*, 2011). It is highly suitable for the improvement of polygenic traits such as drought and salt tolerance because it enables selection for traits controlled by many genes/QTL, with small and large effects (Grenier *et al.*, 2015). Thus, GS captures the effects of many QTL, irrespective of the QTL location in the genome by using linkage disequilibrium (LD). The prediction models are designed

using data on phenotype and genotype in a target breeding population and an individual's selection depends on the breeding values (Spindel and Iwata, 2018). In this approach, a training population is both phenotyped and genotyped to predict the breeding value of the non-phenotyped breeding population. First, a statistical model is calibrated using the training population, then the performances for various traits of the breeding population are predicted using allelic identity with loci that were found to be associated with the phenotype in the training population (Spindel and Iwata, 2018). The genomic estimated breeding values (GEBVs) for lines in the breeding population are estimated based on the model, which is used to predict the performance in the field. GS can be used for prediction of genetic values in each successive generation in a variety of populations (Auinger *et al*., 2016). Several statistical prediction models are available for genomic selection and the accuracy of the model depends on the characters of the population used. Since GS is a new tool for rice improvement, research is needed to develop best-prediction models for improving abiotic stress tolerance (Onogi *et al*., 2015).

Transcriptomics

Transcriptome profiling allows the investigation of a plant's response to abiotic stresses and to identify genes associated with it. At the beginning of the omics era, sequencing of expressed sequence tags (ESTs) from cDNA libraries constructed from different tissues under multiple experimental conditions was used extensively for the discovery of differentially expressed genes in response to different abiotic stresses (Vij and Tyagi, 2007). Later, microarray technique was developed to analyse the global transcriptional changes. It has been used to investigate responses to abiotic stresses (Walia *et al*., 2005, 2007; Degenkolbe *et al*., 2009; Liu *et al*., 2013). A study by Walia *et al*. (2007) at the reproductive stage revealed induction of a large number of genes at panicle initiation stage in salt-sensitive *indica* and *japonica* cultivars compared to salt-tolerant varieties. Many of these salt-responsive genes are ion homeostasis-related genes. A transcription profiling study indicated that more genes were significantly downregulated in the sensitive than in the tolerant cultivars under drought stress, but downregulation of genes involved in photosynthesis was greater in the tolerant than in the sensitive cultivars (Degenkolbe *et al*., 2009).

RNA-Seq strategy is now evolving as a robust and popular tool for investigating genome-wide gene expression and cataloguing of all putative transcripts. It can provide information about alternative gene spliced transcripts, chimeric transcripts, post-transcriptional modifications, changes in expression of a gene over time, or differences in gene expression in different groups or treatments and mutations/SNPs (Maher *et al*., 2009). It involves conversion of a pool of RNA to a cDNA library with adaptors attached to one or both ends, followed by sequencing in a high-throughput manner (Wang *et al*., 2009). The short sequences or reads are either aligned to a reference genome or assembled *de novo* to generate a transcriptome map on a global scale along with their level of expression for each gene.

Moumeni *et al*. (2015) reported potential drought-tolerance pathways and mechanisms through transcriptomic analysis of leaves of NILs and suggested involvement of a transcriptional factor gene '*YABBY*'. An RNA-seq study in a drought-tolerant introgression line (IL) along with the donor and the recurrent parent suggested that genotype-specific drought-induced genes and the genes with higher levels of expression in a drought-tolerant donor under both normal and drought stress conferred drought tolerance (Huang *et al*., 2014). Patil *et al*. (2017) integrated RNA-seq and QTL mapping and identified a wound-inducible protein (*LOC_Os08g08090*) as a candidate gene for drought tolerance in the QTL region. Using a similar approach, Wang *et al*. (2017) identified ten QTL and four candidate genes associated with salinity tolerance at seedling stage in recombinant inbred line (RIL) population involving a *Oryza rufipogon* accession and cultivar 93-11. González-Schain *et al*. (2016) reported repression of many transcription factor genes, signal transduction and metabolic pathway genes in heat-tolerant variety N22. It was further noted that the expression of protective chaperone in anthers was needed to overcome heat-induced damage and facilitate fertilization.

A transcriptome study of cold tolerance at the seedling stage showed fewer differentially expressed genes in the tolerant variety 'Oro' than the susceptible variety 'Tio Taka' and these included genes involved in signal transduction, phytohormones, antioxidant system and biotic stress tolerance (da Maia *et al.*, 2017). Comparison of the transcriptome of a cold-tolerant IL with its recurrent parent revealed that many differentially expressed genes were co-localized on introgressed segments associated with cold tolerance QTL (Zhang *et al.*, 2012). A comparative transcriptome analysis using a deep-water variety C9285 and a non-deep-water variety Taichung 65 revealed a significant difference in expression of genes involved in gibberellin and trehalose biosynthesis, anaerobic fermentation, and cell wall modification as well as expression of ERF genes (Minami *et al.*, 2018).

Proteomics

Proteomics studies are as important as transcriptomics, and their integration can lead to a comprehensive understanding of abiotic stress response because exposure to abiotic stresses result in modification, interaction, movement, *de novo* synthesis and degradation of proteins. It involves application of technologies for the identification and quantification of protein content in a cell, tissue and organism (Ahmad *et al.*, 2016). The rice proteome analysis data are helpful to breeders understanding of the growth and defence mechanisms of the plant. Most common proteomics analysis methods are: 2-D gel electrophoresis, MALDI-TOF-MS, Isotope-coded affinity tags (ICAT), Isobaric Tags for Relative and Absolute Quantification (iTRAQ) and Absolute Quantification (AQUA). Several reviews have been devoted to the utility of a proteomics approach for improving abiotic stress tolerance (Ahmad *et al.*, 2016; Ghatak *et al.*, 2017).

The application of a comparative proteomics approach to organelles and tissues provides information about the amount and quality of proteins, and specific protein modifications in response to abiotic stresses. The rice proteome has been analysed in response to abiotic stresses (reviewed by Subudhi, 2011). Comparative proteomic analysis indicated that proteins involved in abiotic stress tolerance are produced in a larger amount in the early stages of salt-stress in Pokkali compared to salt-susceptible IR64 (Lakra *et al.*, 2018, 2019). Drought proteomic studies were conducted in different tissues of rice plants (Shu *et al.*, 2011; Rabello *et al.*, 2014; Agrawal *et al.*, 2016; Wu *et al.*, 2016). A proteomics study revealed that the drought-tolerant genotype 'Moroberekan' showed better recovery than the drought-sensitive IR64 during anther development after exposure to drought stress (Liu and Bennett, 2011). Mu *et al.* (2017) showed that ribosomal proteins were degraded in response to heat stress in sensitive rice cultivars whereas HSPs, expansins and lipid transfer proteins were increased in resistant cultivars. The above studies indicated stress-responsive proteins as new targets for genome manipulation to improve abiotic stress tolerance in rice.

Metabolomics

Metabolites play a vital role in plant's physiological processes, and their monitoring and quantification could improve abiotic stress tolerance in rice by providing valuable insights into stress-tolerance mechanisms (Fincher *et al.*, 2006). Metabolomics uses the advanced techniques of analytical chemistry and bioinformatics to detect and determine the levels of metabolites using analytical instruments, and various data processing and mining procedures (Oikawa *et al.*, 2008). Assessment of genotypic or phenotypic differences between plant species or among genotypes exhibiting variable tolerance to abiotic stresses could be done using metabolomics data (Pérez-Clemente *et al.*, 2013).

The plant's response to multiple abiotic stresses has been analysed in a number of metabolomic studies (Shulaeva *et al.*, 2008). The metabolomic analysis can be utilized to develop metabolite biomarkers to make genetic progress to improve abiotic stress tolerance. Degenkolbe *et al.* (2013) identified both metabolic and transcript marker candidates for drought tolerance selection. Nitrogen-rich metabolites (amino acids and the nucleotide-related metabolites allantoin and uridine) were accumulated in shoots of the tolerant varieties (Dular and N22) whereas glycolysis and the TCA cycle-related metabolites such as malate, glyceric acid, and glyceric acid-3-phosphate were reduced drastically in

the roots of sensitive genotypes (IR64, IR74) (Casartelli *et al*., 2018). Based on metabolite profiles of FL478 and IR64 under salt stress, Zhao *et al*. (2014) noted that the salt tolerance response of FL478 at early stages was due to a reduction in organic acids, whereas metabolites produced at later stages acted as osmoprotectants. Based on the carbohydrate profiling results and measurements of oxidative products and antioxidative enzymes, Morsy *et al*. (2007) suggested that a more effective reactive oxygen species (ROS) scavenging system might be responsible for chilling tolerance. The utility of metabolomics for genetic analysis of natural variation in abiotic stress tolerance to identify functional genes and metabolic pathways associated with abiotic stress tolerance can be enhanced by integrating it with other omics technologies.

Phenomics

High-throughput phenotyping or phenomics is progressing rapidly to bridge the gap between trait expression and information encoded in the genome. Precise phenotyping is required to assess thousands of lines for crop breeding. Various screening methods have been employed in rice to measure tolerance to abiotic stresses. However, most of the protocols to measure plant biomass are destructive, which poses a limitation for the measurement of active abiotic stress responses in plant growth (Das *et al*., 2015). To address this obstacle, the use of imaging technologies has increased in plant science research (Jansen *et al*., 2014). Various non-destructive phenotyping techniques include magnetic resonance imaging (MRI), thermal imaging and chlorophyll fluorescence. MRI can be used to measure water content and water transport in different plant tissues non-invasively (Van As *et al*., 2009). Infrared thermography is a new phenotypic method to detect genetic variation in the stomatal response to water deficit in a controlled environment as well as in the field (James *et al*., 2010). Image capturing and robotic technologies have increased the precision and speed of phenotyping. Automated plant phenotyping technology provides a precise tool for the characterization of plant stress responses (Singh *et al*., 2018). Integration of genomic selection and high-throughput phenotyping can allow evaluation of large populations to make rapid gains in abiotic stress tolerance (Juliana *et al*., 2019). The field of phenomics will enable the discovery of new traits and sources of new genes at a lower cost than traditional phenotypic approaches.

Transgenic approach

Transgenic strategies to improve abiotic stress tolerance initially targeted genes involved in synthesis of osmoprotectants, detoxifying enzymes, dehydrins, molecular chaperones, transport proteins and water channel proteins. Since the single-gene approach was not sufficient, signalling and pathways regulating expression of a large number of downstream genes were exploited. Numerous genetic engineering studies have been conducted to improve abiotic stress tolerance in rice (reviewed by Karan and Subudhi, 2011; Manju Latha *et al*., 2017; Nguyen *et al*., 2018). A large number of abiotic stress-responsive genes are available to develop abiotic stress-tolerant rice varieties (Kumar *et al*., 2013; Reddy *et al*. 2017; Li *et al*., 2018).

Overexpression of *OsNHX1* from 'Pokkali' altered Na and K accumulation in rice root and shoot and improved germination and biomass production under salt stress (Amin *et al*., 2016). Similarly, overexpression of *PtCYP714A3*, *SUV3* and Pea DNA helicase (*PDH45*) conferred salt tolerance in rice by maintaining photosynthesis and antioxidant machinery and reducing Na^+ concentration in the shoot (Nath *et al*., 2016; Wang *et al*., 2016). A Na^+/K^+ transporter gene of rice, *OsSOS1*, enhanced salinity tolerance in transgenic BRRI Dhan 28 by decreasing Na^+ content (Yasmin *et al*., 2016). Rice plants overexpressing *OsPP1a* showed a higher level of salt tolerance, which could be due to upregulation of salt-induced genes like *OsNAC5*, *OsRK1A* and *OsNAC6* (Liao *et al*., 2016). Biradar *et al*. (2018) developed transgenic rice lines with salt responsive protein 3-1 (*SaSRP3-1*) gene as well as pyramided lines with *SaSRP3-1* and vacuolar H^+-ATPase subunit c1 (*SaVHAc1*) derived from a halophyte grass *Spartina alterniflora*. Evaluation of these transgenic lines revealed that both single-gene and pyramided plants had enhanced salt tolerance at the seedling and vegetative stages, but pyramids performed better at the reproductive stages. Transgenic rice plants with overexpression

of isoforms of glutamine synthetase *OsGS1;1* and *OsGS2* improved both osmotic and salt tolerance at the seedling stage (James *et al.*, 2018).

Numerous recent reports have also highlighted enhanced drought tolerance in transgenic rice plants using *OsAHL1* (Zhou *et al.*, 2017), *OsHSP50.2* (Xiang *et al.*, 2018), *OsJAZ1* (Fu *et al.*, 2017) and *OsASR5* (Jinjie *et al.*, 2017). Transgenic rice plants overexpressing transcription factors have been shown to enhance both drought and salinity tolerance (Xiong *et al.*, 2014; Tang *et al.*, 2019). Although this approach provides a rapid way to generate abiotic stress-tolerant plants, the progress has been confined to laboratory environments because of consumer resistance to genetically modified rice.

Genome editing

Precise targeted changes in the genome by genome editing have become a powerful tool in molecular biology (Adli, 2018). It involves the use of site-specific nucleases (SSNs), such as zinc finger nucleases (ZFNs), transcriptional activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR)-associated endonuclease Cas9 (CRISPR/Cas9). These nucleases make double-strand breaks (DSBs) in the target DNA, which are subsequently repaired by non-homologous end joining (NHEJ) or homologous recombination (HR). In most cases, random insertions or deletions caused by NHEJ result in gene knockout due to frameshift mutations in the coding region of a gene. On the other hand, precise gene modifications or gene insertions can be accomplished by the homologous recombination-mediated repair mechanism (Bortesi and Fischer, 2015). Among these tools, ZFNs and TALENs are not popular because of low efficiency, high cost and technical complexity. The CRISPR/Cas9 system is emerging as a simple, efficient and cost-effective tool for genome engineering in plants. It utilizes a chimeric single guide RNA, which directs Cas9 endonuclease to make a DSB in a target region of the genome, followed by DNA repair through the NHEJ or HR pathway. It is possible to target any location of the genome by choosing a short guide RNA.

Genome editing can bring about a revolutionary change in improving abiotic stress tolerance at a much faster rate through precise genome engineering. Rice is an ideal model organism for genome editing technology because of its small genome size and abundant genetic resources (Mishra *et al.*, 2018). Although there are many examples of genome editing in rice for other characteristics like yield, quality and disease resistance, few reports are available on abiotic stress tolerance. The functional role of *OsDERF1, OsPMS3, OsPSPS, OsMSH1, OsEPSPS, OsMSH1* and *OsMYB5* under drought stress was identified via the CRISPR/Cas9 system (Zhang *et al.*, 2014). Lou *et al.* (2017) successfully used CRISPR/Cas9 to produce a *SAPK2* rice mutant that conferred drought tolerance. Shi *et al.* (2016) created novel variants of a gene *ARGOS8*, a negative regulator of ethylene responses, which improved grain yield under drought stress in maize. In *Arabidopsis*, a mutant of a gene for a proton pump, *OST2*, generated by this tool, changed the stomatal closing pattern (Osakabe *et al.* 2016). A transcription factor gene (*OsNAC041*) was mutated by CRISPR-Cas9 and the rice mutant exhibited susceptibility to salt (Wang *et al.*, 2019), whereas there was increased salt tolerance in the mutant for the *OsRR22* gene (Zhang *et al.*, 2019). Some potential future uses of CRISPR technology are multiplexing or simultaneous manipulation of multiple genes, construction of genome-wide mutant libraries, and regulation of gene expression.

Conclusion and Future Prospects

Climate change is affecting food security in almost every corner of the world. Prolonged drought, extreme temperature, submergence and salinity pose major threats to rice production. Since the growing world population will require 60% more food by 2050, development of new rice varieties with tolerance to multiple abiotic stresses is an urgent need to enhance food production. This can be accomplished by utilizing natural genetic variability and coupling novel omics tools with traditional approaches (Mickelbart *et al.*, 2015). Promising methods are being developed to discover and functionally characterize novel genes controlling abiotic stress tolerance (Takeda and Matsuoka, 2008; Varshney *et al.*, 2014). The integration of genetic modification strategies with

marker-assisted pyramiding of stress-tolerance QTL/genes and genomic selection along with high-throughput phenotyping tools will accelerate progress towards development of high-yielding abiotic stress-tolerant rice varieties (Ashikari and Matsuoka, 2006; Takeda and Matsuoka, 2008). CRISPR/Cas9-mediated genome manipulation is now emerging as a cost-effective, precise and rapid next-generation breeding method to create novel genetic variation and design rice varieties with desirable abiotic stress-tolerance traits (Ye and Cui, 2019). Efforts should be made to manipulate the transcription factor and transporter genes to enhance abiotic stress tolerance without compromising yield (Mickelbart *et al*., 2015). The utilization of genome-editing tools for crop improvement will be accelerated with the discovery of superior allelic variants and enhanced understanding of the gene network involved in abiotic stress tolerance.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., *et al*. (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology* 30, 174–178.

Adli, M. (2018) The CRISPR tool kit for genome editing and beyond. *Nature Communications* 9, 1911.

Agrawal, L., Gupta, S., Mishra, S.K., Pandey, G., Kumar, S., *et al*. (2016) Elucidation of complex nature of PEG induced drought-stress response in rice root using comparative proteomics approach. *Frontiers in Plant Science* 7, 1466.

Ahmad, P., Latef, A.A.H.A., Rasool, S., Akram, N.A., Ashraf, M., *et al*. (2016) Role of proteomics in crop stress tolerance. *Frontiers in Plant Science* 7, 1336.

Ahmadi, N., Negrão, S., Katsantonis, D., Frouin, J., Ploux, J., *et al*. (2011) Targeted association analysis identified *japonica* rice varieties achieving Na (+)/K (+) homeostasis without the allelic make-up of the salt tolerant *indica* variety Nona Bokra. *Theoretical and Applied Genetics* 123, 881–895.

Ali, J., Xu, J.L., Gao, Y.M., Ma, X.F., Meng L.J., *et al*. (2017) Harnessing the hidden genetic diversity for improving multiple abiotic stress tolerance in rice (*Oryza sativa* L.). *PLoS ONE* 12, e0172515.

Al-Tamimi, N., Brien, C., Oakey, H., Berger, B., Saade, S., *et al*. (2016). Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nature Communications* 7, 13342.

Amin, U.S.M., Biswas, S., Elias, S.M., Razzaque, S., Haque, T., *et al*. (2016) Enhanced salt tolerance conferred by the complete 2.3 Kb cDNA of the rice vacuolar Na(+)/H(+) antiporter gene compared to 1.9 kb coding region with 5′ UTR in transgenic lines of rice. *Frontiers in Plant Science* 7, 14.

Andaya, V.C. and Mackill, D.J. (2003) Mapping of QTLs associated with cold tolerance during the vegetative stage in rice. *Journal of Experimental Botany* 54, 2579–2585.

Anyaoha, C.O., Fofana, M., Gracen, V., Tongoona, P. and Mande, S. (2019) Introgression of two drought QTLs into FUNAABOR-2 early generation backcross progenies under drought stress at reproductive stage. *Rice Science* 26, 32–41.

Ashikari, M. and Matsuoka, M. (2006) Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends in Plant Science* 11, 344–350.

Asoro, F.G., Newell, M.A., Beavis, W.D., Scott, M.P. and Jannink, J.L. (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4, 132.

Auinger, H.-J., Schönleben, M., Lehermeier, C., Schmidt, M. , Korzun, V., *et al*. (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theoretical and Applied Genetics* 129, 2043–2053.

Bhattarai, U. and Subudhi, P.K. (2018a) Genetic analysis of yield and agronomic traits under reproductive stage drought stress in rice using a high-resolution linkage map. *Gene* 669, 69–76.

Bhattarai, U. and Subudhi, P.K. (2018b) Identification of drought responsive QTLs during vegetative growth stage of rice using a saturated GBS-based SNP linkage map. *Euphytica* 214, 38.

Biradar, H., Karan, R. and Subudhi, P.K. (2018) Overexpression of a salt responsive protein3-1 as well as pyramiding with *SaVHAc1* from *Spartina alterniflora* L. enhances salt tolerance in rice. *Frontiers in Plant Science* 9, 1304.

Bizimana, J.B., Luzi, K., Murori, R.L. and Singh, R.K. (2017) Identification of quantitative trait loci for salinity tolerance in rice (*Oryza sativa* L.) using IR29/Hasawi mapping population. *Journal of Genetics* 96, 571–582.

Blumwald, E. (2000) Sodium transport and salt tolerance in plants. *Current Opinion in Cell Biology* 12, 431–434.

Bonilla, P., Dvorak, J., Mackill, D.J., Deal, K. and Gregorio, G. (2002) RLFP and SSLP mapping of salinity tolerance genes in chromosome 1 of rice (*Oryza sativa* L.) using recombinant inbred lines. *Philippine Journal of Agricultural Science* 85, 68–76.

Bortesi, L. and Fischer, R. (2015) The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances* 33, 41–52.

Boyer, J.S. (1982) Plant productivity and environment. *Science* 218, 443–448.

Casartelli, A., Riewe, D., Hubberten, H.M., Altmann, T., Hoefgen, R., *et al*. (2018) Exploring traditional aus-type rice for metabolites conferring drought tolerance. *Rice* 11, 9.

Catolos, M., Sandhu, N., Dixit, S., Shamsudin, N.A.A., Naredo, M.E.B., *et al*. (2017) Genetic loci governing grain yield and root development under variable rice cultivation conditions. *Frontiers in Plant Science* 8, 1763.

da Maia, L.C., Cadore, P.R.B., Benitez, L.C., Danielowski, R., Braga, E.J.B., *et al*. (2017) Transcriptome profiling of rice seedlings under cold stress. *Functional Plant Biology* 44, 419–429.

Dar, M.H., Zaidi, N.W., Waza, S.A., Verulkar, S.B., Ahmed, T., *et al*. (2018) No yield penalty under favorable conditions paving the way for successful adoption of flood tolerant rice. *Scientific Reports* 8, 9245.

Das, G. and Rao, G.J.N. (2015) Molecular marker assisted gene stacking for biotic and abiotic stress resistance genes in an elite rice cultivar. *Frontiers in Plant Science* 6, 698.

Das, G., Patra, J.K. and Baek, K.H. (2017) Insight into MAS: A molecular tool for development of stress resistant and quality of rice through gene stacking. *Frontiers in Plant Science* 8, 985.

Das, P., Nutan, K.K., Singla-Pareek, S.L. and Pareek, A. (2015) Understanding salinity responses and adopting 'omics-based' approaches to generate salinity tolerant cultivars of rice. *Frontiers in Plant Science* 6, 712.

De Leon, T.B., Linscombe, S., Gregorio, G. and Subudhi, P.K. (2015) Genetic variation in Southern USA rice genotypes for seedling salinity tolerance. *Frontiers in Plant Science* 6, 374.

De Leon, T.B., Linscombe, S., and Subudhi, P.K. (2016) Molecular dissection of seedling salinity tolerance in rice (*Oryza sativa* L.) using a high-density GBS-based SNP linkage map. *Rice* 9, 52.

De Leon, T.B., Linscombe, S., and Subudhi, P.K. (2017) Identification and validation of QTLs for seedling salinity tolerance in introgression lines of a salt tolerant rice landrace 'Pokkali'. *PLoS ONE* 12, e0175361.

Degenkolbe, T., Do, P.T., Zuther, E., Repsilber, D., Walther, D., *et al*. (2009) Expression profiling of rice cultivars differing in their tolerance to long-term drought stress. *Plant Molecular Biology* 69, 133–153.

Degenkolbe, T., Do, P.T., Kopka, J., Zuther, E., Hincha, D.K., *et al*. (2013) Identification of drought tolerance markers in a diverse population of rice cultivars by expression and metabolite profiling. *PLoS ONE* 8, e63637.

Dharmappa, P.M., Doddaraju, P., Malagondanahalli, M.V., Rangappa, R.B., Mallikarjuna, N.M., *et al*. (2019) Introgression of root and water use efficiency traits enhances water productivity: An evidence for physiological breeding in rice (*Oryza sativa* L.). *Rice* 12, 14.

Endo, T., Chiba, B., Wagatsuma, K., Saeki, K., Ando, T., *et al*. (2016) Detection of QTLs for cold tolerance of rice cultivar 'Kuchum' and effect of QTL pyramiding. *Theoretical and Applied Genetics* 129, 631–640.

Fincher, G., Paltridge, N. and Langridge, P. (2006) Functional genomics of abiotic stress tolerance in cereals. *Briefings in Functional Genomics* 4, 343–354.

Fu, J., Wu, H., Ma, S.Q., Xiang, D.H., Liu, R.Y., *et al*. (2017) OsJAZ1 attenuates drought resistance by regulating JA and ABA signaling in rice. *Frontiers in Plant Science* 8, 13.

Fukao, T. and Xiong, L. (2013) Genetic mechanisms conferring adaptation to submergence and drought in rice: Simple or complex? *Current Opinion in Cell Biology* 16, 196–204.

Ghatak, A., Chaturvedi, P. and Weckwerth, W. (2017) Cereal crop proteomics: Systemic analysis of crop drought stress responses towards marker-assisted selection breeding. *Frontiers in Plant Science* 8, 757.

Gimhani, D.R., Gregorio, G.B. and Kottearachchi, N.S. (2016) SNP-based discovery of salinity-tolerant QTLs in a bi-parental population of rice (*Oryza sativa*). *Molecular Genetics Genomics* 291, 2081–2099.

González-Schain, N., Dreni, L., Lawas, L.M.F., Galbiati, M., Colombo, L., *et al*. (2016) Genome-wide transcriptome analysis during anthesis reveals new insights into the molecular basis of heat stress responses in tolerant and sensitive rice varieties. *Plant Cell Physiology* 57, 57–68.

Gregorio, G.B., Senadhira, D., Mendoza, R.D., Manigbas, N.L., Roxas, J.P., *et al*. (2002) Progress in breeding for salinity tolerance and associated abiotic stresses in rice. *Field Crop Research* 76, 91–101.

Grenier, C., Cao, T.V., Ospina, Y., Quintero, C., Châtel, M.H., *et al*. (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PLoS ONE* 10, e0154976.

Grondin, A., Dixit, S., Torres, R., Venkateshwarlu, C., Rogers, E., *et al*. (2018) Physiological mechanisms contributing to the QTL qDTY3.2 effects on improved performance of rice Moroberekan x Swarna $BC_2F_{3:4}$ lines under drought. *Rice* 11, 43.

Hallajian, M.T. (2016) Mutation breeding and drought stress tolerance in plants. In: Hossain, M.A., Wani, S.H., Bhattacharjee, S., Burritt, D.J., and Tran, L.S.P. (eds) *Drought Stress Tolerance in Plants, Volume 2*. Springer, Cham, Switzerland, pp. 359–383.

Hattori, Y., Nagai, K., Furukawa, S., Song, X.J., Kawano, R., Sakakibara, H., *et al*. (2009) The ethylene response factors *SNORKEL1* and *SNORKEL2* allow rice to adapt to deep-water. *Nature* 460, 1026–1030.

Hay, S., Oo, M., Minn, M., Linn, K.Z., Mar, N.N., *et al*. (2015) Development of drought tolerant mutant from rice var. Manawthukha through mutation breeding technique using 60 Co Gamma source. *International Journal of Innovative Research in Science, Engineering and Technology* 4, 11205–11212.

Hoang, G.T., Van Dinh, L., Nguyen, T.T., Ta, N.K., Gathignol, F., *et al*. (2019) Genome-wide association study of a panel of Vietnamese rice landraces reveals new QTLs for tolerance to water deficit during the vegetative phase. *Rice* 12, 4.

Howarth, C.J. and Ougham, H.J. (1993) Gene expression under temperature stress. *New Phytologist* 125, 1–26.

Huang, L., Zhang F., Zhang, F., Wang, W., Zhou, Y., *et al*. (2014) Comparative transcriptome sequencing of tolerant rice introgression line and its parents in response to drought stress. *BMC Genomics* 15, 1026.

Hussain, S., Zhang, J.H., Zhong, C., Zhu, L.F., Cao, X.C., *et al*. (2017) Effects of salt stress on rice growth, development characteristics, and the regulating ways: A review. *Journal of Integrative Agriculture* 16, 2357–2374.

Islam, S., Ontoy, J. and Subudhi, P.K. (2019) Meta-analysis of quantitative trait loci (QTLs) associated with seedling stage salt tolerance in rice (*Oryza sativa* L.). *Plants* 8, 33.

Ismail, A.M. (2018) Submergence tolerance in rice: Resolving a pervasive quandary. *New Phytologist* 218, 1298–1300.

Jagadish, S.V.K., Muthurajan, R., Oane, R., Wheeler, T.R., Heuer, S., *et al*. (2010) Physiological and proteomic approaches to address heat tolerance during anthesis in rice (*Oryza sativa* L.). *Journal of Experimental Botany* 61, 143–156.

James, D., Borphukan, B., Fartyal, D., Ram, B., Singh, J., *et al*. (2018) Concurrent overexpression of OsGS1;1 and OsGS2 genes in transgenic rice (*Oryza sativa* L.): Impact on tolerance to abiotic stresses. *Frontiers in Plant Science* 9, 786.

James, R.A., Munns, R., Furbank, R.T., Sirault, X.R.R. and Jones, H.G. (2010) New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *Journal of Experimental Botany* 61, 3499–3507.

Jansen, M., Pinto, F., Nagel, K.A., van Dusschoten, D., Fiorani, F., *et al*. (2014) Non-invasive phenotyping methodologies enable the accurate characterization of growth and performance of shoots and roots. In: Tuberosa, R., Graner, A. and Frison, E. (eds) *Genomics of Plant Genetic Resources: Managing, Sequencing and Mining Genetic Resources*. Springer, Dordrecht, The Netherlands, pp. 173–206.

Ji, K., Wang, Y., Sun, W., Lou, Q., Mei, H., *et al*. (2012) Drought-responsive mechanisms in rice genotypes with contrasting drought tolerance during reproductive stage. *Journal of Plant Physiology* 169, 336–344.

Jinjie, L., Yang, L., Zhigang, Y., Jihong, J., Minghui, Z., *et al* (2017) *OsASR5* enhances drought tolerance through a stomatal closure pathway associated with ABA and H_2O_2 signalling in rice. *Plant Biotechnology Journal* 15, 183–196.

Juliana, P., Montesinos-López O.A., Crossa, J., Mondal, S., González Pérez, L., *et al*. (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theoretical and Applied Genetics* 132, 177–194.

Karan, R. and Subudhi, P.K. (2011) Approaches to increasing salt tolerance in crop plants. In: Ahmad, P. and Prasad, M.N.V. (eds), *Abiotic Stress Responses in Plants: Metabolism to Productivity*. Springer, New York, pp. 63–88.

Kilasi, N.L., Singh, J., Vallejos, C.E., Ye, C., Jagadish, S.V.K., *et al*. (2018) Heat stress tolerance in rice (*Oryza sativa* L.): Identification of under heat stress quantitative trait loci and candidate genes for seedling growth. *Frontiers in Plant Science* 9, 1578.

Kretzschmar, T., Pelayo, M.A.F., Trijatmiko, K.R., Gabunada, L.F., Alam, R., *et al*. (2015) A trehalose-6-phosphate phosphatase enhances anaerobic germination tolerance in rice. *Nature Plants* 1, 15124.

Kumar, A., Dixit, S., Ram, T., Yadaw, R.B., Mishra, K.K., *et al*. (2014) Breeding high-yielding drought-tolerant rice: Genetic variations and conventional and molecular approaches. *Journal of Experimental Botany* 65, 6265–6278.

Kumar, A., Sandhu, N., Dixit, S., Yadav, S., Swamy, B.P.M., *et al*. (2018) Marker-assisted selection strategy to pyramid two or more QTLs for quantitative trait-grain yield under drought. *Rice* 11, 35.

Kumar, K., Kumar, M., Kim, S.R., Ryu, H. and Cho, Y.G. (2013) Insights into genomics of salt stress response in rice. *Rice* 6, 27.

Kumar, V., Singh, A., Mithra, S.V.A., Krishnamurthy, S.L., Parida, S.K., *et al*. (2015) Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Research* 22, 133–145.

Kurokawa, Y., Nagai, K., Huan, P.D., Shimazaki, K., Qu, H., *et al*. (2018) Rice leaf hydrophobicity and gas films are conferred by a wax synthesis gene (LGF1) and contribute to flood tolerance. *New Phytologist* 218, 1558–1569.

Lakra, N., Kaur, C., Anwar, K., Singla-Pareek, S.L. and Pareek, A. (2018) Proteomics of contrasting rice genotypes: Identification of potential targets for raising crops for saline environment. *Plant Cell and Environment* 41, 947–969.

Lakra, N., Kaur, C., Singla-Pareek, S.L. and Pareek, A. (2019) Mapping the 'early salinity response' triggered proteome adaptation in contrasting rice genotypes using iTRAQ approach. *Rice* 12, 3.

Lekklar, C., Pongpanich, M., Suriya-arunroj, D., Chinpongpanich, A., Tsai, H., *et al*. (2019) Genome-wide association study for salinity tolerance at the flowering stage in a panel of rice accessions from Thailand. *BMC Genomics* 20, 76.

Li, L. and Mao, D. (2018) Deployment of cold tolerance loci from *Oryza sativa* ssp. *Japonica* cv. 'Nipponbare' in a high-yielding *Indica* rice cultivar '93-11'. *Plant Breeding* 137, 553–560.

Li, Y., Xiao, J., Chen, L., Huang, X., Cheng, Z., *et al*. (2018) Rice functional genomics research: Past decades and future. *Molecular Plant* 11, 359–380.

Liao, Y.D., Lin, K.H., Chen, C.C. and Chiang, C.M. (2016) *Oryza sativa* protein phosphatase 1a (OsPP1a) involved in salt stress tolerance in transgenic rice. *Molecular Breeding* 36, 22.

Liu, C., Ou, S., Mao, B., Tang, J., Wang, W., *et al*. (2018) Early selection of *bZIP73* facilitated adaptation of *japonica* rice to cold climates. *Nature Communication* 9, 3302.

Liu, F., Xu, W., Song, Q., Tan, L., Liu, J., *et al*. (2013) Microarray-assisted fine-mapping of quantitative trait loci for cold tolerance in rice. *Molecular Plant* 6, 757–767.

Liu, J.X. and Bennett, J. (2011) Reversible and irreversible drought-induced changes in the anther proteome of rice (*Oryza sativa* L.) genotypes IR64 and Moroberekan. *Molecular Plant* 4, 59–69.

Liu, Q., Yang, T., Yu, T., Zhang, S., Mao, X., *et al*. (2017) Integrating small RNA sequencing with QTL mapping for identification of miRNAs and their target genes associated with heat tolerance at the flowering stage in rice. *Frontiers in Plant Science* 8, 43.

Lou, D., Wang, H., Liang, G. and Yu. D. (2017) *OsSAPK2* confers abscisic acid sensitivity and tolerance to drought stress in rice. *Frontiers in Plant Science* 8, 993.

Ma, X., Feng, F., Wei, H., Mei, H., Xu, K., *et al*. (2016) Genome-wide association study for plant height and grain yield in rice under contrasting moisture regimes. *Frontiers in Plant Science* 7, 1801.

Ma, Y., Dai, X., Xu, Y., Luo, W., Zheng, X., *et al*. (2015) *COLD1* confers chilling tolerance in rice. *Cell* 160, 1209–1221.

Maas, E.V. (1990) Crop salt tolerance. In: Tanji, K.K. (ed.) *Agricultural Salinity Assessment and Management*. American Society of Civil Engineers, New York, pp. 262–304.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., *et al*. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458, 97–101.

Manju Latha, G., Mohapatra, T., Swapna Geetanjali, A. and Sambasiva Rao, K.R.S. (2017) Engineering rice for abiotic stress tolerance: A review. *Current Trends in Biotechnology and Pharmacy* 11, 396–413.

Mao, D., Xin, Y., Tan, Y., Hu, X., Bai, J., *et al*. (2019) Natural variation in the *HAN1* gene confers chilling tolerance in rice and allowed adaptation to a temperate climate. *Proceedings of the National Academy of Sciences USA* 116, 3494.

Mickelbart, M.V., Hasegawa, P.M., Bailey-Serres, J. (2015) Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nature Review Genetics* 16, 237–251.

Minami, A., Yano, K., Gamuyao, R., Nagai, K., Kuroha, T., *et al*. (2018) Time-course transcriptomics analysis reveals key responses of submerged deep water rice to flooding. *Plant Physiology* 176, 3081–3102.

Mishra, R., Joshi, R.K. and Zhao K. (2018) Genome editing in rice: Recent advances, challenges, and future implications. *Frontiers in Plant Science* 9, 1361.

Mohammadi, R., Mendioro, M.S., Diaz, G.Q., Gregeorio, G.B. and Singh, R.K. (2013) Mapping quantitative trait loci associated with yield and yield components under reproductive stage salinity stress in rice (*Oryza sativa* L.). *Journal of Genetics* 92, 433–443.

Morsy, M.R., Jouve, L., Hausman, J.F., Hoffmann, L. and Stewart, J.M. (2007) Alteration of oxidative and carbohydrate metabolism under abiotic stress in two rice (*Oryza sativa* L.) genotypes contrasting in chilling tolerance. *Journal of Plant Physiology* 164, 157–167.

Moumeni, A., Satoh, K., Venuprasad, R., Serraj, R., Kumar, A., *et al*. (2015) Transcriptional profiling of the leaves of near-isogenic rice lines with contrasting drought tolerance at the reproductive stage in response to water deficit. *BMC Genomics* 16, 1110.

Mu, Q.L., Zhang, W.Y., Zhang, Y.B., Yan, H.L., Liu, K., *et al*. (2017) iTRAQ-based quantitative proteomics analysis on rice anther responding to high temperature. *International Journal of Molecular Sciences* 18, 1811.

Munns, R. and Tester, M. (2008) Mechanisms of salinity tolerance. *Annual Review of Plant Biology* 59, 651–681.

Nath, M., Yadav, S., Kumar Sahoo, R., Passricha, N., Tuteja, R., *et al*. (2016) PDH45 transgenic rice maintain cell viability through lower accumulation of Na+, ROS and calcium homeostasis in roots under salinity stress. *Journal of Plant Physiology* 191, 1–11.

Negrão, S., Courtois, B., Ahmadi, N., Abreu, I., Saibo, N., *et al*. (2011) Recent updates on salinity stress in rice: from physiological to molecular responses. *Critical Reviews in Plant Science* 30, 329–377.

Nguyen, H.C., Lin, K.H., Ho, S.L., Chiang, C.M. and Yang, C.M. (2018) Enhancing the abiotic stress tolerance of plants: From chemical treatment to biotechnological approaches. *Physiologia Plantarum* 164, 452–466.

Nogoy, F.M., Song, J.Y., Ouk, S., Rahimi, S., Kwon, S.W., *et al*. (2016) Current applicable DNA markers for marker assisted breeding in abiotic and biotic stress tolerance in rice (*Oryza sativa* L.). *Plant Breeding Biotechnology* 4, 271–284.

Oikawa, A., Matsuda, F., Kusano, M., Okazaki, Y. and Saito, K. (2008) Rice metabolomics. *Rice* 1, 63–71.

Onogi, A., Ideta, O., Inoshita, Y., Ebana, K., Yoshioka, T., *et al*. (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* 128, 41–53.

Örvar, B.L., Sangwan, V., Omann, F. and Dhindsa, R.S. (2000) Early steps in cold sensing by plant cells: The role of actin cytoskeleton and membrane fluidity. *The Plant Journal* 23, 785–794.

Osakabe, Y., Watanabe, T., Sugano, S.S., Ueta, R., Ishihara, R., *et al*. (2016) Optimization of CRISPR/Cas9 genome editing to modify abiotic stress responses in plants. *Scientific Reports* 6, 26685.

Pan, Y., Zhang, H., Zhang, D., Li, J. and Xiong, H. (2015) Genetic analysis of cold tolerance at the germination and booting stages in rice by association mapping. *PLoS ONE* 10, e0120590.

Pandit, E., Tasleem, S., Barik, S.R., Mohanty, D.P. and Nayak, D.K. (2017) Genome-wide association mapping reveals multiple QTLs governing tolerance response for seedling stage chilling stress in *indica* rice. *Frontiers in Plant Science* 8, 552.

Pantalião, G., Narciso, M., Guimarães, C., Castro, A., Colombari, J., *et al*. (2016) Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica* 144, 651–664.

Patil, S., Srividhya, A., Veeraghattapu, R., Deborah, D.A.K., Kadambari, G.M., *et al*. (2017) Molecular dissection of a genomic region governing root traits associated with drought tolerance employing a combinatorial approach of QTL mapping and RNA-seq in rice. *Plant Molecular Biology Reporter* 35, 457–468.

Pérez-Clemente, R.M., Vives, V., Zandalinas, S.I., López-Climent, M.F., Muñoz, V., *et al*. (2013) Biotechnological approaches to study plant responses to stress. *BioMedical Research International* 2013, 654120. DOI: 10.1155/2013/654120.

Prince, S.J., Beena, R., Gomez, S.M., Senthivel, S. and Babu, R.C. (2015) Mapping consistent rice (*Oryza sativa* L.) yield QTLs under drought stress in target rainfed environments. *Rice* 8, 25.

Punyawaew, K., Suriya-arunroj, D., Siangliw, M., Thida, M., Lanceras-Siangliw, J., *et al*. (2016) Thai jasmine rice cultivar KDML105 carrying *Saltol* QTL exhibiting salinity tolerance at seedling stage. *Molecular Breeding* 36, 150.

Puram, V.R.R., Ontoy, J., Linscombe, S. and Subudhi, P.K. (2017) Genetic dissection of seedling stage salinity tolerance in rice using introgression lines of a salt tolerant landrace Nona Bokra. *Journal of Heredity* 108, 658–670.

Puram, V.R.R., Ontoy, J. and Subudhi, P.K. (2018) Identification of QTLs for salt tolerance traits and prebreeding lines with enhanced salt tolerance using a salt tolerant donor 'Nona Bokra'. *Plant Molecular Biology Reporter* 36, 695–709.

Rabara, R., Msanne, J., Ferrer, M. and Basu, S. (2018) When water runs dry and temperature heats up: Understanding the mechanisms in rice tolerance to drought and high temperature stress. *Preprints* 2018060426.

Rabello, F.R., Villeth, G.R., Rabello, A.R., Rangel, P.H., Guimarães, C.M., *et al*. (2014) Proteomic analysis of upland rice (*Oryza sativa* L.) exposed to intermittent water deficit. *The Protein Journal* 33, 221–230.

Rahman, M.A., Bimpong, I.K., Bizimana, J.B., Pascual, E.D., Arceta, M., *et al*. (2017). Mapping QTLs using a novel source of salinity tolerance from Hasawi and their interaction with environments in rice. *Rice* 10, 47.

Reddy, I.N.B.L., Kim, B.K., Yoon, I.S., Kim, K.H. and Kwon, T.R. (2017) Salt tolerance in rice: Focus on mechanisms and approaches. *Rice Science* 24, 123–144.

Ren, Z.H., Gao, J.P., Li, L.G., Cai, X.L., Huang, W., *et al*. (2005) A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nature Genetics* 37, 1141.

Saleem, M.Y., Mukhtar, Z., Cheema, A.A. and Atta, B.M. (2005) Induced mutation and in vitro techniques as a method to induce salt tolerance in Basmati rice (*Oryza sativa* L.). *International Journal of Environmental Science and Technology* 2, 141–145.

Sandhu, N., Dixit, S., Swamy, B.P.M., Raman, A., Kumar, S., *et al*. (2019) Marker assisted breeding to develop multiple stress tolerant varieties for flood and drought prone areas. *Rice* 12, 8.

Schläppi, M.R., Jackson, A.K., Eizenga, G.C., Wang, A., Chu, C., *et al*. (2017) Assessment of five chilling tolerance traits and GWAS mapping in rice using the USDA mini-core collection. *Frontiers in Plant Science* 8, 957.

Shah, F., Nie, L., Cui, K., Shah, T., Wu, W., *et al*. (2014) Rice grain yield and component responses to near 2°C of warming. *Field Crop Research* 157, 98–110.

Shamsudin, N.A., Swamy, B.P.M., Ratnam, W., Cruz, M.T.S., Sandhu, N., *et al*. (2016) Pyramiding of drought yield QTLs into a high quality Malaysian rice cultivar MRQ74 improves yield under reproductive stage drought. *Rice* 9, 21.

Shanmugavadivel, P.S., Mithra, A.S., Prakash, C., Ramkumar, M.K., Tiwari, R., *et al*. (2017). High resolution mapping of QTLs for heat tolerance in rice using a 5K SNP array. *Rice* 10, 28.

Shi, J., Gao, H., Wang, H., Lafitte, H.R., Archibald, R.L., *et al*. (2016) ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnology Journal* 15, 207–216.

Shu, L., Lou, Q., Ma, C., Ding, W., Zhou, J., *et al*. (2011) Genetic, proteomic and metabolic analysis of the regulation of energy storage in rice seedlings in response to drought. *Proteomics* 11, 4122–4138.

Shulaeva, V., Cortesa, D., Miller, G. and Mittler, R. (2008) Metabolomics for plant stress response. *Physiologia Plantarum* 132, 199–208.

Singh, A., Septiningsih, E.M., Balyan, H.S., Singh, N.K. and Rai, V. (2017) Genetics, physiological mechanisms and breeding of flood-tolerant rice (*Oryza sativa* L.). *Plant Cell Physiology* 58, 185–197.

Singh, B., Mishra, S., Bohra, A., Joshi, R. and Siddique. K.H.M. (2018) Crop phenomics for abiotic stress tolerance in crop plants. In: Wani, S.H. (ed.) *Physiological and Molecular Avenues for Combating Abiotic Stress Tolerance in Plants*, Elsevier Academic Press, London, pp. 277–296.

Singh, R., Singh, Y., Xalaxo, Y., Verulkar, S., Yadav, V., *et al* (2016) From QTL to variety-harnessing the benefits of QTLs for drought, flood and salt tolerance in mega rice varieties of India through a multi-institutional network. *Plant Science* 241, 278–287.

Spindel, J. and Iwata, H. (2018) Genomic selection in rice breeding In: Sasaki, T. and Ashikari, M. (eds) *Rice Genomics, Genetics and Breeding*. Springer Singapore, Singapore, pp. 473–496.

Subudhi, P.K. (2011) Omics approaches for abiotic stress tolerance in plants. In: Tuteja, N., Gill, S.S. and Tuteja, R. (eds) *E-book: Omics and Plant Abiotic Stress Tolerance*. Bentham Science Publishers Ltd, Sharjah, United Arab Emirates. DOI: 10.2174/97816080505811110101, pp. 10–38.

Sun, J., Yang, L., Wang, J., Liu, H., Zheng, H., *et al*. (2018) Identification of a cold-tolerant locus in rice (*Oryza sativa* L.) using bulked segregant analysis with a next-generation sequencing strategy. *Rice* 11, 24.

Swamy, B.P.M., Shamsudin, N.A.A., Rahman, S.N.A., Mauleon, R., Ratnam, W., *et al*. (2017) Association mapping of yield and yield related traits under reproductive stage drought stress in rice (*Oryza sativa* L.). *Rice* 10, 21.

Takagi, H., Tamiru, M., Abe, A., Yoshida, K., Uemura, A., *et al*. (2015) MutMap accelerates breeding of a salt-tolerant rice cultivar. *Nature Biotechnology* 33, 445–449.

Takeda, S. and Matsuoka, M. (2008) Genetic approaches to crop improvement: Responding to environmental and population changes. *Nature Review Genetics* 9, 444–457.

Tang, Y.H., Bao, X.X., Zhi, Y.L., Wu, Q., Guo, Y.R., *et al*. (2019) Overexpression of a MYB family gene, *OsMYB6*, increases drought and salinity stress tolerance in transgenic rice. *Frontiers in Plant Science* 10, 12.

Tiwari, S., Krishnamurthy, S.L., Kumar, V., Singh, B., Rao, A.R., *et al*. (2016) Mapping of QTLs for salt tolerance in rice (*Oryza sativa* L.) by bulked segregant analysis of recombinant inbred lines using 50K SNP chip. *PLoS ONE* 11, e0153610.

Uga, Y., Sugimoto, K., Ogawa, S., Tane, J., Ishitani, M., *et al*. (2013) Control of root system architecture by deeper rooting 1 increases rice yield under drought conditions. *Nature Genetics* 45, 1097–1102.

Valarmathi, M., Sasikala, R., Rahman, R., Jagadeeshselvam, R., Rohit, K., *et al*. (2019) Development of salinity tolerant version of a popular rice variety improved white ponni through marker assisted back cross breeding. *Plant Physiology Reports* 24(2), 262–271. DOI: 10.1007/s40502-019-0440-x.

Van As, H., Scheenen, T. and Vergeldt, F.J. (2009) MRI of intact plants. *Photosynthesis Research* 102, 213–222.

Varshney, R.K., Terauchi, R. and McCouch, S.R. (2014) Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biology* 12, e1001883.

Vij, S. and Tyagi, A.K. (2007) Emerging trends in the functional genomics of the abiotic stress response in crop plants. *Plant Biotechnology Journal* 5, 361–380.

Vikram, P., Swamy, B.P.M., Dixit, S., Ahmed, H.U., Cruz, M.T.S., *et al*. (2011) $qDTY_{1.1,}$ a major QTL for rice grain yield under reproductive-stage drought stress with a consistent effect in multiple elite genetic backgrounds. *BMC Genetics* 12, 89.

Walia, H., Wilson, C., Condamine, P., Liu, X., Ismail, A.M., *et al*. (2005) Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. *Plant Physiology* 139, 822–835.

Walia, H., Wilson, C., Zeng, L.H., Ismail, A.M., Condamine, P., *et al*. (2007) Genome-wide transcriptional analysis of salinity stressed *japonica* and *indica* rice genotypes during panicle initiation stage. *Plant Molecular Biology* 63, 609–623.

Wang, B., Zhong Z., Zhang H., Wang, X., Liu, B., *et al*. (2019) Targeted mutagenesis of NAC transcription factor gene, *OsNAC041*, leading to salt sensitivity in rice. *Rice Science* 26, 98–108.

Wang, C., Yang, Y., Wang, H., Ran, X., Li, B., *et al*. (2016) Ectopic expression of a cytochrome P450 monooxygenase gene *PtCYP714A3* from *Populus trichocarpa* reduces shoot growth and improves tolerance to salt stress in transgenic rice. *Plant Biotechnology Journal* 14, 1838–1851.

Wang, S., Cao, M., Ma, X., Chen, W., Zhao, J., *et al*. (2017) Integrated RNA sequencing and QTL mapping to identify candidate genes from *Oryza rufipogon* associated with salt tolerance at the seedling stage. *Frontiers in Plant Science* 8, 1427.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10, 57–63.

Wu, Y.Q., Mirzaei, M., Pascovici, D., Chick, J.M., Atwell, B.J., *et al*. (2016) Quantitative proteomic analysis of two different rice varieties reveals that drought tolerance is correlated with reduced abundance of photosynthetic machinery and increased abundance of ClpD1 protease. *Journal of Proteomics* 143, 73–82.

Xiang, J., Chen, X., Hu, W., Xiang, Y., Yan, M., *et al*. (2018) Overexpressing heat-shock protein OsHSP50.2 improves drought tolerance in rice. *Plant Cell Report* 37, 1585–1595.

Xiao, N., Gao, H., Qian, H., Gao, Q., Wu, Y., *et al*. (2018) Identification of genes related to cold tolerance and a functional allele that confers cold tolerance. *Plant Physiology* 177, 1108–1123.

Xiong, H., Li, J., Liu, P., Duan, J., Zhao, Y., *et al*. (2014) Overexpression of OsMYB48-1, a novel MYB-related transcription factor, enhances drought and salinity tolerance in rice. *PLoS ONE* 9, e92913.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., *et al*. (2006) *Sub1A* is an ethylene-response factor-like gene that confers submergence tolerance to rice. *Nature* 442, 705–708.

Yasmin, F., Biswas, S., Jewel, G.M., Elias, S. and Seraj, Z. (2016) Constitutive overexpression of the plasma membrane Na^+/H^+ antiporter for conferring salinity tolerance in rice. *Plant Tissue Culture and Biotechnology* 25, 257–272.

Ye, C., Tenorio, F.A., Redona, E.D., Morales-Cortezano, P.S., Cabrega, G.A., *et al*. (2015). Fine-mapping and validating *qHTSF4.1* to increase spikelet fertility under heat stress at flowering in rice. *Theoretical and Applied Genetics* 128, 1507–1517.

Ye, J. and Cui, X. (2019) Next-generation crop breeding methods. *Molecular Plant* 12, 470–471.

Zhang, A., Liu, Y., Wang, F., Li, T., Chen, Z., *et al*. (2019) Enhanced rice salinity tolerance via CRISPR/Cas9-targeted mutagenesis of the *OsRR22* gene. *Molecular Breeding* 39, 47.

Zhang, F., Huang, L., Wang, W., Zhao, X., Zhu, L., *et al*. (2012) Genome-wide gene expression profiling of introgressed *indica* rice alleles associated with seedling cold tolerance improvement in a *japonica* rice background. *BMC Genomics* 13, 461.

Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., *et al*. (2014). The CRISPR/Cas9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnology Journal* 12, 797–807.

Zhang, J., Zheng, H.G., Aarti, A., Pantuwan, G., Nguyen, T.T., *et al*. (2001a) Locating genomic regions associated with components of drought resistance in rice: Comparative mapping within and across species. *Theoretical and Applied Genetics* 103, 19–29.

Zhang, M., Lu, Q., Wu, W., Niu, X., Wang, C., *et al*. (2017) Association mapping reveals novel genetic loci contributing to flooding tolerance during germination in *indica* rice. *Frontiers in Plant Science* 8, 678.

Zhang, Q., Chen, Q., Wang, S., Hong, Y. and Wang Z. (2014) Rice and cold stress: Methods for its evaluation and summary of cold tolerance-related quantitative trait loci. *Rice* 7, 24.

Zhang, W.P., Shen, X.Y., Wu, P., Hu, B. and Liao CY (2001b) QTLs and epistasis for seminal root length under a different water supply in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* 103, 118–123.

Zhang, Z., Li, J., Pan, Y., Li, J., Zhou, L., *et al*. (2017) Natural variation in *CTB4a* enhances rice adaptation to cold habitats. *Nature Communication* 8, 14788.

Zhao, J., Zhang, S., Dong, J., Yang, T., Mao, X., *et al*. (2017) A novel functional gene associated with cold tolerance at the seedling stage in rice. *Plant Biotechnology Journal* 15, 1141–1148.

Zhao, X., Wang, W., Zhang, F., Deng, J., Li, Z., *et al*. (2014) Comparative metabolite profiling of two rice genotypes with contrasting salt stress tolerance at the seedling stage. *PLoS ONE* 9, e108020.

Zhou, L., Liu, Z., Liu, Y., Kong, D., Li, T., *et al*. (2017) A novel gene *OsAHL1* improves both drought avoidance and drought tolerance in rice. *Scientific Report* 6, 30264.

Zhu, S., Huang, R., Wai, H.P., Xiong, H., Shen, X., *et al*. (2017) Mapping quantitative trait loci for heat tolerance at the booting stage using chromosomal segment substitution lines in rice. *Physiology and Molecular Biology of Plants* 23, 817–825.

17 Quantitative Genetics, Molecular Techniques and Agronomic Performance of Provitamin A Maize in Sub-Saharan Africa

Baffour Badu-Apraku[1]*, M.A.B. Fakorede[2], A.O. Talabi[1], E. Obeng-Bio[3], S.G.N. Tchala[4] and S.A. Oyekale[5]

[1]*International Institute of Tropical Agriculture, Ibadan, Nigeria;* [2]*Obafemi-Awolowo University, Ile-Ife, Osun State, Nigeria;* [3]*CSIR-Crops Research Institute, Fumesua, Kumasi, Ghana;* [4]*Institut Togolais de Recherche Agronomique (ITRA), Lomé Cacaveli,Togo;* [5]*Ladoke Akintola University of Technology, Ogbomoso, Nigeria*

Introduction

Maize (*Zea mays* L.), a crop introduced to sub-Saharan Africa (SSA) more than 500 years ago, has become a major staple food crop in the region. It is the most widely consumed food staple in Africa, providing approximately 30% of the total calories to more than 4.5 billion people in developing countries. Poor-quality diets, often deficient in minerals and vitamins, dominate much of SSA. Vitamin A is one of the vitamins deficient in the diets, which is important for proper immune-system function. Approximately one-third of children under the age of 5 years are at risk of vitamin A deficiency (VAD), the leading cause of childhood blindness. Results of some studies have shown that 15.3% of pregnant women in SSA are deficient in vitamin A (Aguayo and Baker, 2005; World Health Organization [WHO], 2009). This is because the normal endosperm maize that is commonly grown and used as staple food in the sub-region is nutritionally deficient in provitamin A (PVA) (Safawo *et al.*, 2010; Venado *et al.*; 2017) and two essential amino acids – lysine and tryptophan – which, unfortunately, the human body cannot synthesize (Sofi *et al.*, 2009; Nuss and Tanumihardjo, 2011; Le *et al.*, 2016). This predisposes millions of people, who live mainly on maize as a staple food, to VAD and protein deficiency. In addition to causing night blindness (nyctalopia), VAD increases childhood mortality, and results in growth retardation and depressed immune response (West and Darnton-Hill, 2008; WHO, 2009; Muthayya *et al.*, 2013), while kwashiorkor and pellagra are the negative effects of protein deficiency and low levels of tryptophan in the body.

Results of nutrition trials, in which synthesized vitamin A was administered, showed, on average, a 24% reduction in child mortality. Therefore, biofortification of maize has been identified as an economical and sustainable strategy for tackling PVA deficiency (Sagare *et al.*, 2015a) in maize, in SSA. The HarvestPlus Challenge Programme has established 15 μg g^{-1} as the breeding target for PVA maize hybrids and

* Email: b.badu-apraku@cgiar.org

© CAB International, 2020.

open-pollinated varieties (OPVs). However, only a few released PVA maize hybrids have attained this level in SSA. Therefore, efforts are underway to develop and commercialize varieties with elevated levels of PVA. Quality protein maize (QPM) has been proposed as a possible food source for ameliorating protein malnutrition not only in SSA but also in other parts of the world (Nuss and Tanumihardjo, 2011). Although efforts have been made to develop varieties of maize that can mitigate the effects of VAD and protein malnutrition (Fan *et al.*, 2004; Krivanek *et al.*, 2007; Sofi *et al.*, 2009; Azmach *et al.*, 2013; Suwarno *et al.*, 2014; Badu-Apraku *et al.*, 2015a,b; Menkir *et al.*, 2017; Tandzi *et al.*, 2017; Gebremeskel *et al.*, 2018), maize varieties with combined PVA and quality protein characteristics that will simultaneously solve the problems of VAD and protein deficiency are yet to be developed and commercialized in SSA.

Furthermore, maize production in SSA is constrained by myriad abiotic and biotic stresses, some of which are infestation by *Striga hermonthica* (giant witchweed), terminal drought and dry spells at any growth stage of the maize crop, low soil nutrients, especially nitrogen, diseases and insect pests. Development and deployment of QPM varieties with multiple stress tolerance and high levels of PVA that will simultaneously solve the problems of malnutrition and food insecurity in SSA have been among the major goals of the International Institute of Tropical Agriculture's (IITA) maize improvement programme. The maize germplasm available to the breeders contains many inbred lines, OPVs and hybrids, into which tolerance to some or all of the stresses and/or quality proteins have been incorporated. The materials available from our germplasm, along with others obtained from sources such as the International Maize and Wheat Improvement Center (CIMMYT), were screened for PVA, subjected to quantitative genetic studies and molecular approaches for PVA enhancement. Through this programme, multiple stress-tolerant OPVs, inbred lines and hybrids with high-quality protein and PVA levels have been developed for release to farmers of SSA. Described briefly in this chapter are the materials and methods used, the results obtained, and the OPVs and hybrids that have been released and those in the pipeline for release and commercialization in SSA.

Quantitative Genetics of Maize Provitamin A

For a long time, the existence of PVA components and/or precursors in maize was not known about. Theoretically, if PVA had been a qualitative trait, it would have been detected much earlier than when it was identified. Maize breeders hypothesized that it was a quantitative trait, but its mode of inheritance, genetic diversity and distribution, heritability, response to selection, phenotypic and genetic correlations with other maize traits, and several other necessary pieces of genetic information needed for its thorough understanding and improvement, were unknown. Quantitative traits are controlled by multiple genes, each segregating according to Mendel's laws. They are also affected by the environment to varying degrees, thus making their predictability from the phenotypic measurements rather difficult and unreliable. For this reason, breeders conduct extensive studies on the quantitative aspects of traits of the crops used in their research. For PVA, the maize germplasm has been screened extensively, its genetic diversity/variation for PVA content investigated, the mode of inheritance and heritability determined, the genotype × environment interaction quantified and its response to selection for population improvement initiated.

Screening maize germplasm collection at the International Institute of Tropical Agriculture for provitamin A

The PVA maize breeding programme is led by the CIMMYT and IITA, in collaboration with public- and private-sector research partners in SSA, and is supported by the HarvestPlus Challenge Programme (www.harvestplus.org). In 2012, PVA breeding efforts resulted in the release of three maize hybrids in Zambia and two in Nigeria, with total PVA carotenoid concentrations of more than 7 μg g^{-1}, and experimental cultivars with 10–15 μg g^{-1} have been identified (Saltzman *et al.*, 2013; Dhliwayo *et al.*, 2014; Suwarno *et al.*, 2014). The global germplasm banks of the CGIAR institutes and the germplasm banks held in trust by national partners provide a reservoir of germplasm of staple crops for screening by HarvestPlus. Genetic transformation provides an alternative strategy to

incorporate specific genes that express nutritional density.

The first step in conventional breeding is to determine whether sufficient genetic variation exists to breed for a particular trait of interest, which, in the specific case of HarvestPlus, included sourcing of parental populations with target or higher levels of iron, zinc and PVA. Researchers have analysed approximately 300,000 maize samples for trace minerals or for PVA carotenoids during screening (Menkir *et al*., 2008). Some studies have reported significant genetic variation for carotenoids in yellow maize lines and hybrids adapted to temperate environments (Brunson and Quackenbush, 1962; Grogan *et al*., 1963; Quackenbush *et al*., 1966; Weber, 1987; Kurilich and Juvik, 1999).

Genetic diversity/variation for provitamin A content in the maize germplasm

Crop improvement activities focus first on exploring the available genetic diversity for Fe, Zn and PVA carotenoids. At the same time or during subsequent screening, agronomic and end-use features are characterized. The objectives when exploring the available genetic diversity are to identify (i) parental genotypes that can be used in crosses, genetic studies, molecular marker development and parent-building; and (ii) existing varieties, pre-varieties in the release pipeline or finished germplasm products for 'fast-tracking'. Fast-tracking means releasing, commercializing or introducing genotypes that combine the target micronutrient density with the required agronomic and end-user traits so they can be quickly delivered to producers and have immediate impact on micronutrient-deficient populations (Pfeiffer and McClafferty, 2007).

A source of genetic variation is essential for the next breeding steps. If variation is present in the strategic gene pool, pre-breeding is required. If variation is present in the tactical gene pool, the materials may be used directly to develop competitive varieties. Most breeding programmes simultaneously conduct pre-breeding and final product enhancement to develop germplasm combining high levels of one or more micronutrients. If the available genetic variation suggests that target micronutrient increments are unlikely to be reached, it is still possible to find genetic variation through transgressive segregation or by exploiting heterosis. When variation is not available, a transgenic approach may be the only remaining option (Bouis *et al*., 2002; Khush, 2002; Al-Babili and Beyer, 2005).

Breeding for increased concentrations of PVA is promising because there is considerable genetic variation for PVA available in maize germplasm. Studies initially conducted by CIMMYT revealed that among 1000 tropical maize genotypes, total PVA varied from 0.24 to 8.80 μg g^{-1}, while the proportion of PVA to total carotenoids ranged between 5 and 30% (Ortiz-Monasterio *et al*., 2007). Furthermore, the HarvestPlus project has been conducting extensive work on improving PVA levels in elite maize lines, hybrids and synthetic populations.

To date in Africa, more than 40 PVA maize synthetics, single-cross hybrids and three-way hybrids have been released in the Democratic Republic of the Congo (DRC), Ghana, Malawi, Mali, Nigeria, Rwanda, Tanzania, Zambia and Zimbabwe. The first wave of varieties released in 2012/13 contained 6–8 ppm (6–8 μg g^{-1}) additional PVA (about 50% of the target increment), while the second-wave of varieties (released in 2015/16) contained about 10 ppm (10 μg g^{-1}) additional PVA (66% of the target increment) (Pixley *et al*., 2013). Varieties that fully meet the PVA target level are being tested in multi-location trials across SSA and are expected to be released by 2020. All biofortified varieties combine competitive grain yield and consumer-preferred end-use quality traits with increased PVA content. Additional crop improvement research is underway to develop PVA maize with enhanced carotenoid stability, to reduce the rate and pace of carotenoid degradation in storage and end-use (Ortiz *et al*., 2016).

Maize exhibits considerable natural variation for kernel carotenoids, with some lines accumulating as much as 66 μg g^{-1}. The predominant carotenoids in maize kernels, in decreasing order of concentration, are lutein, zeaxanthin, β-carotene, β-cryptoxanthin and α-carotene. Among lines included in the IITA diverse maize panel, PVA levels have reached 23.98 μg g^{-1}. However, most yellow maize grown and consumed throughout the world has only 0.5 to 1.5 μg g^{-1} PVA levels. Generally, yellow/orange maize grains possess lower levels of PVA as compared to non-PVA carotenoids.

Carotenoids in maize kernels produce colours ranging from light yellow to dark orange, and they are concentrated primarily in the vitreous (horny) portion of the endosperm (Weber, 1987). Darker orange kernels in maize tend to have higher total carotenoid concentrations compared to lighter orange and yellow kernels (Harjes *et al.*, 2008), but most of the increased total concentration results from non-PVA xanthophyll carotenoids (Quackenbush *et al.*, 1961; Weber, 1987). Even though considerable variation exists for total carotenoid levels in maize kernels, Harjes *et al.* (2008) showed that the majority of yellow kernel maize inbred lines from a global collection did not have sufficient PVA (average of 1.7 μg g^{-1} β-carotene) to meet recommended dietary allowance levels for adequate nutrition (US Institute of Medicine, 2001).

All yellow genotypes of maize contain carotenoids, although the fraction of carotenoids with PVA activity (β-cryptoxanthin, α- and β-carotene, which can be converted to vitamin A) is typically small (e.g. 10–20%) compared with zeaxanthin and lutein (each around 30–50% of total carotenoids) (Brenna and Berardo, 2004; Howe and Tanumihardjo, 2006). β-carotene and β-cryptoxanthin are the most abundant PVAs in maize, whereas α-carotene is present in much smaller amounts. There is considerable variation, however, in the ratios of total PVA to total carotenoid concentrations, as well as in the ratio of β-carotene to β-cryptoxanthin. Given the considerable natural variation identified to date for total PVA concentration (about 0–15 mg g^{-1} at HarvestPlus and 0 to almost 9 mg g^{-1} at CIMMYT), and in view of the wide range in ratios among carotenoids in maize, we concluded that there is considerable scope for breeding maize with enhanced PVA concentration and improved nutritional value.

Analyses of genotypes with yellow to dark orange kernels have identified large variation in the number of PVA molecules (Egesel *et al.*, 2003a) and their carotenoid profiles. At CIMMYT, carotenoid profiles of more than 1000 tropical maize genotypes have been analysed and promising materials have been identified with PVA concentrations (~8 mg g^{-1}) and/or carotenoid profiles that could be used in breeding programmes to increase total PVA content in the grain. To date, there has been no consistent trend in the relationship between geographical origin of maize genotypes and the highest PVA concentrations; the best materials include pale-yellow temperate, dark-yellow highland tropical and intense-orange lowland tropical lines (Ortiz-Monasterio *et al.*, 2007). Although a substantial breeding effort is needed, genetic variation appears adequate to achieve nutritionally significant concentrations of PVA carotenoids in maize grain (Ortiz-Monasterio *et al.*, 2007).

Maize germplasm resources exhibit wide genetic diversity (Liu *et al.*, 2003) with corresponding variation in carotenoid profiles (Harjes *et al.*, 2008), features that are useful for investigating pathway regulation and generating breeding populations. The close evolutionary relationship between maize and other food crops in the Poaceae family provides an opportunity for exploitation of gene homologues in other grass species for improvement of PVA levels in maize through genome synteny. Considerable diversity exists in the regulation of synthesis and catabolism of carotenoids (Auldridge *et al.*, 2006; Vallabhaneni *et al.*, 2010; Arango *et al.*, 2014). Studies of carotenoid content and composition in maize grains have identified significant allelic variation for key genes, such as *lycopene epsilon cyclase* (*LCYE*) (Harjes *et al.*, 2008) and *β-carotene hydroxylase1* (*CRTRB1*) (Yan *et al.*, 2010), which govern critical steps in the pathway, leading to the successful use of marker-assisted selection (MAS) in applied breeding programmes (Babu *et al.*, 2013).

The carotenoid association mapping (CAM) panel consisted of 380 diverse lowland tropical (47%), subtropical (47%) and temperate (3%) lines assembled by CIMMYT's HarvestPlus-funded maize breeding programme. The panel includes ten lines in which a PVA-enhancing allele of *CRTRB1* has been incorporated through MAS (Babu *et al.*, 2013). Initial screening of more than 1500 maize germplasm accessions revealed ranges for PVA from 0–19 ppm (0–19 μg g^{-1}) in existing maize varieties, exceeding the PVA target of 15 ppm (15 μg g^{-1}) (Ortiz-Monasterio *et al.*, 2007; Harjes *et al.*, 2008). These nutrients were consistently expressed in the maize inbred lines across different growing conditions, and further assessment indicated potential to increase the levels of multiple carotenoids simultaneously (Dhliwayo *et al.*, 2014; Suwarno *et al.*, 2014). The identification of loci associated with PVA carotenoids and the development of DNA

markers have led to accelerated genetic gains in breeding for increased PVA content. The most important PVA-enhancing alleles identified to date are lycopene epsilon cyclase (*lcyE*) and β-carotene hydroxylase 1 (*crtRB1*) (Yan *et al*., 2010; Suwarno *et al*., 2014). Validation experiments showed that the latter alone often doubles, and sometimes triples, the total concentration of PVA carotenoid content in maize grain, mainly by increasing the β-carotene content (Babu *et al*., 2013). The PVA maize breeding programmes at CIMMYT, IITA and the Zambia Agriculture Research Institute (ZARI) began in 2007. The breeding pipeline includes materials from the two lead institutions, CIMMYT (tropical mid-altitude) and IITA (tropical lowlands), as well as local germplasm. Both hybrid and open-pollinated (synthetic) biofortified varieties are being developed.

Laboratory screening methods

Given the large number of materials to be analysed and the short turnaround time for doing sample analysis of crops with two or more cycles per year, breeding effectively for minerals and PVA carotenoids depends on the availability of low-cost and quick high-throughput screening methods (HTMs) (Pfeiffer and McClafferty, 2007). Rapid techniques for screening cereals, legumes and tubers for minerals and PVA are currently being developed, validated and implemented at various Consultative Group for International Agricultural Research (CGIAR) centres and national research institutes. These research efforts include developing protocols for conventional analytical methods, given that sample preparation, as well as digestion, extraction and milling procedures need to be standardized, and that the accuracy of participating laboratories must be assessed by external quality assurance programmes (Pfeiffer and McClafferty, 2007).

The simplest screening approach for carotenoid concentration in maize is to visually select dark orange or yellow seed. However, the correlation between visually assessed colour and PVA content is small. Using a HunterLab miniscan (CIELab scale for colour assessment), it was found that chroma measured on 15% hydrated flour was correlated to PVA content ($r = 0.58$, $P < 0.05$), indicating that the miniscan may be effective as a preliminary selection tool (Lozano-Alejo *et al*., 2007). Sensitive, accurate biochemical methods have been used for carotenoid quantification in grain or processed food. High performance liquid chromatography (HPLC) methods, with a diode array detector useful for carotenoid determinations, have been reported (Weber, 1987; Kurilich and Juvik, 1999; Gama *et al*., 2005). When using HPLC, difficulties may arise during carotenoid extraction, given that carotenoids are very sensitive to heat, acids, light and/or oxygen; this may lead to structural changes and quantification errors. Consequently, the extraction procedure can be time consuming, with only about 30 samples analysed per day. Less complicated than HPLC is inductively coupled plasma-optical emission spectrophotometry (ICP-OES), an accurate methodology for microelement quantification. However, the high cost (US$55 per sample in the case of HPLC and U$5.00–7.50 for ICP-OES) and low throughput of these techniques make them inadequate for use in breeding programmes.

Carotenoid analyses were conducted at CIMMYT's maize quality laboratory in Mexico. Random samples of 50 seeds were kept frozen at −80°C until ground to a fine powder (0.5 μm), followed by the use of the CIMMYT laboratory protocols for carotenoid analysis that included extraction, separation and quantification by HPLC for TL10 and TL11 environments (Galicia *et al*., 2008), and by ultra-performance liquid chromatography (UPLC) for AF12 (Galicia *et al*., 2012). The procedure followed in the analysis has been described in detail by Babu *et al*. (2013). Briefly, 50 kernels per entry were used for the carotenoid analysis via HPLC. Carotenoids were released from finely ground dried maize grain samples by adding ethanol. Samples were then saponified, followed by carotenoid extraction using hexane. Carotenoid separation and quantification were done using HPLC with a C30 column attached to a C30 filter insert. A multi-wavelength detector set at 450 nm was used, and data were collected and processed using Waters Millennium 2010 software (Waters Chromatography). Lutein, zeaxanthin, β-cryptoxanthin, and all-trans-β-carotene were identified through their characteristic spectra and

comparison of their retention times with known standard solutions. Total PVA content (μg g^{-1}) was calculated for each sample as the sum of β-carotene plus one-half of β-cryptoxanthin. The HarvestPlus project has investigated the application of near infrared reflectance spectroscopy (NIRS) for rapid and inexpensive semi-quantitative screening of maize samples for PVA (Ortiz-Monasterio *et al.*, 2007).

Discovery efforts to understand key genes involved in natural variation for carotenoid content have used genome-wide association study (GWAS) approaches to explore allelic variations at loci previously established to be associated with the carotenoid pathway in maize or other model species (Harjes *et al.*, 2008; Vallabhaneni *et al.*, 2009; Yan *et al.*, 2010). With the availability of high-density genotyping platforms, such as Illumina's infinium (MaizeSNP50) and genotyping by sequencing (GBS) (Elshire *et al.*, 2011), it is now possible to quickly generate millions of marker data points that are distributed throughout the genome. The GWAS based on high density, extensive marker coverage increases our ability to explain the inheritance of target traits (Gibson, 2010; Stranger *et al.*, 2011).

Genetic variability for provitamin A and its components

Genetic variability is fundamental to improvement of any economic traits, such as carotenoids. Yellow maize is the only grain crop that naturally accumulates a significant amount of carotenoids in its seed (Buckner *et al.*, 1990). There is tremendous variation in concentrations of PVA in yellow maize, resulting in a pronounced genetic variability in germplasm collections, which can be visualized as cream, butter, yellow or orange endosperm. Significant genetic variations for carotenoid content has been reported for temperate maize (Forgey, 1974; Egesel *et al.*, 2003b; Mishra and Singh, 2010). Blessin *et al.* (1963) reported ranges of 0.9 to 4.1 μg g^{-1} for carotenes and 18.6 to 48.0 μg g^{-1} for xanthophylls for 39 maize inbred lines. Quackenbush *et al.* (1963) also reported PVA contents ranging from a trace to 7.3 μg g^{-1}, and lutein from 2 to 33 μg g^{-1} for 125 inbred lines. In another study involving a diverse panel of inbred lines, β-carotene level was found to be up to 13.6 μg g^{-1}, whereas most of the yellow maize grown and consumed throughout the world has only 0.5 to 1.5 μg g^{-1} β-carotene (Harjes *et al.*, 2008). Mishra and Singh (2010) also reported total carotenoid contents to vary from a minimum of 0.027 μg g^{-1} to a maximum of 25.75 μg g^{-1} dry weight, with an overall mean of 18.11 μg g^{-1} dry weight in a diverse panel of maize inbred lines.

Similarly, among tropical maize germplasm, studies have established wide genetic variability for PVA. The tropical adapted yellow maize varieties grown in Africa naturally contain some levels of PVA and non-PVA carotenoids (xanthophylls), with potential health benefits to humans, more importantly, for eye health (Muzhingi *et al.*, 2008). Though the adapted yellow maize contains an appreciable amount (2 μg g^{-1}) of PVA, it is still considered insufficient and therefore, it cannot meet a significant proportion of daily human requirements (FAO, 1994).

Muthusamy *et al.* (2015) in a multi-location study involving 105 maize inbreds of indigenous and exotic origin revealed wide genetic variation for lutein (0.2–11.3 μg g^{-1}), zeaxanthin (0.2–20.0 μg g^{-1}) and β-carotene (0.0–15.0 μg g^{-1}). For β-cryptoxanthin, variation observed was low (0.1–3.3 μg g^{-1}). Carotenoids were quite stable across environments that played a minor role in causing variation. Islam (2004) reported wide variation in carotenoid profiles in maize, even among inbred lines with similar estimated dietary PVA activity (expressed as vitamin A or retinol activity) in addition to variation for PVA concentrations per se. Similarly, the PVA fraction of total carotenoids varies widely (5–30%) among CIMMYT inbred lines evaluated to date (N. Palacios-Rojas, unpublished results). This suggests that a breeding approach in which inbred lines are chosen as parents for hybrids based on complementary PVA profiles may be successful.

Based on findings from temperate maize germplasm, the dominant carotenoids in maize kernels, in decreasing order of concentration, are: lutein, zeaxanthin, β-carotene, β-cryptoxanthin and α-carotene. Similarly, in tropically adapted maize inbred lines developed during the prebreeding activities of IITA in an effort to biofortify maize, the yellow-endosperm maize genotypes exhibited significant differences for all the traits measured. Seventy per cent of the total variation observed was attributable to β-carotene (Menkir and

Maziya-Dixon, 2004). The mean β-carotene content averaged across six environments varied from 0.45 to 2.18 μg g^{-1}. Multiple stress-tolerant extra-early maize inbreds with PVA levels higher than the target of 15 μg g^{-1} established by the Harvest-Plus Programme have been identified at IITA, e.g. TZEEIOR 202 (23.98 μg g^{-1}) and TZEEIOR 205 (22.58 μg g^{-1}). Furthermore, an early-maturing inbred line, TZEIORQ 55 (15.1 μg g^{-1}), has been identified. The extra-early and early PVA inbreds are invaluable sources of high PVA for developing high PVA hybrids and introgression of PVA alleles into tropical breeding populations. Marker-assisted recurrent selection is being adopted to accumulate favourable PVA alleles in tropical maize breeding populations and elite inbred lines in the IITA maize improvement programme.

Mode of inheritance and heritability estimates

Progress in breeding maize for increased carotenoid content could be rapid, as suggested by the high heritability estimates (Wong *et al*., 1998) and the preponderance of additive genetic variance for carotenoid content in maize (Brunson and Quackenbush, 1962; Grogan *et al*., 1963; Egesel *et al*., 2003a). Egesel *et al*. (2003a), for example, reported that general combining ability (GCA) effects, or additive gene action, accounted for 72–87% of the variation for β-carotene, β-cryptoxanthin and total carotenoids in a diallel study involving ten maize lines. Non-additive gene action was important, however, for PVA concentrations in some crosses (Egesel *et al*., 2003a) and suggested the possibility of exploiting heterosis in breeding for these nutrients. A study by Brunson and Quackenbush (1962) demonstrated clearly that the PVA content of all single-cross hybrids among high PVA inbred lines was, on average, 4.4 times more than that of all single-cross hybrids among lines with low PVA content. Medium to high (0.55–0.90) heritability and a preponderance of additive over non-additive effects determined PVA concentrations in maize (Egesel *et al*., 2003a; Menkir *et al*., 2014; Suwarno *et al*., 2014) which suggested that recurrent selection for PVA content should be effective (Hallauer and Miranda, 1988; Coors, 1999). However, while recurrent selection generally has proven effective for modifying quantitative traits in maize, correlated effects are less predictable (Dhliwayo *et al*., 2014). Muthusamy *et al*. (2015) reported high heritability (>90%) and genetic advance (>75%) for all the carotenoid components. They also reported that zeaxanthin showed a positive correlation with lutein and β-cryptoxanthin, while β-carotene, the major PVA carotenoid, did not show a high correlation with other carotenoids. Kernel colour was positively correlated with lutein (r=0.25), zeaxanthin (r=0.47) and β-cryptoxanthin (r=0.44), but not with β-carotene (r=0.04). This suggested that visual selection based on kernel colour will be misleading in selecting PVA-rich genotypes.

Similar to the findings of Egesel *et al*. (2003a) in a ten-parent diallel, results obtained from a IITA North Carolina Design II (NCD II) study of PVA-QPM inbred lines indicated that the GCA effects were greater than specific combining ability (SCA) effects for all the carotenoids (Obeng-Bio, 2019). The GCA sum of squares ranged from 73% of the genotype sum of squares for total carotenoids to 90% for β-carotene, whereas SCA ranged from 10% for β-carotene to 27% for total carotenoids. The GCA effects accounted for 87% for PVA, whereas the three PVA carotenoid components, β-cryptoxanthin, α-carotene and β-carotene, accounted for GCA effects of 83, 81 and 90%, respectively. It was striking that the GCA-female effects for PVA and β-carotene were relatively larger than the GCA-male effects, whereas the GCA-male effects for β-cryptoxanthin and α-carotene were greater than the GCA-female effects. Halilu (2016), in a half-diallel study, reported significant genotypic differences for lutein and zeaxanthin. In addition, the GCA and SCA variances were significant for lutein and zeaxanthin. The ratio of GCA:SCA of less than unity (<1) was reported for all PVA active carotenoids, with the exception of βCX. These indicated preponderance of non-additive gene action for the measured traits, whereas lutein, zeaxanthin and βCX had ratios equal to or greater than one, indicating a preponderance of additive effects over dominance effects. Narrow-sense heritability estimates across environments ranged from 0.00% for αC to 49.20% for lutein. Chander *et al*. (2008) reported heritability estimates for nutritional traits in maize to be medium (65.6% for protein) to high

(92.5% for R and γ tocopherol ratio). Similar results reported earlier were within this range (Wong *et al.*, 2004). Chander *et al.* (2008) obtained high heritability estimates for carotenoids (84–96%), whereas Wong *et al.* (2004) reported medium to high levels of heritability for fractions of carotenoids (48–87%).

Information on patterns of inheritance of a trait assists in the choice of the most appropriate method for improving a crop for that trait. Thus, heritability estimates are useful in predicting gain from selection and comparing the gain from selection under different experimental designs for devising optimal breeding strategies (Hallauer and Carena, 2009). Inheritance of carotenoid content has been a topical issue in maize (Ford, 2000). In general, carotenoid content is heritable and thus can be improved through plant breeding. The effect of maternal contribution to maize endosperm carotenoid was reported by Egesel *et al.* (2003b) to be supposedly attributable to the diploid contribution of the mother plant to the endosperm, which is triploid. The effect of environment on carotenoids was also found to be small (Egesel *et al.*, 2003a; Menkir and Maziya-Dixon, 2004; Menkir *et al.*, 2008), indicating the possibility of developing maize hybrids with consistently high levels of PVA. The genetic advance as per cent of mean (GAM) for nutritional traits of maize kernels ranged from 6.1% for starch to 22.5% for β-carotene (Chander *et al.*, 2008). The higher GAM for carotenoids and tocopherols suggested the possibility of greater scope of improvements for these nutritionally important compounds. Since heritability estimates must refer to a defined population of genotypes and specified population of environments (Nyquist and Baker, 1991), there is a need to study the heritability of the different carotenoids of maize breeding populations developed by IITA for PVA biofortification activities. This is expected to provide a guide for effective selection and population development.

In a quantitative trait loci (QTL) study, heritability of carotenoid on a line-mean basis across all Nested Associated Mapping (NAM) families ($\hat{h}_1^2$) corresponded to the maximum level of phenotypic variability among lines from the ten NAM families, which could be attributed to the combined effects of QTL (Hung and Holland, 2012). Kernel colour had a moderately high line-mean basis heritability($\hat{h}_1^2 = 0.78 \pm 0.05$) across 2 years at a single location, indicating that enough statistical power and precision existed for QTL mapping and effect estimation. The estimated heritability of kernel colour on an individual plot basis ($h^2 = 0.69 \pm 0.06$), where plot (one-row) was an unreplicated experimental unit that consisted of a single line, was only slightly lower than ˆ2 *l h*. Estimated ˆ2 *lwp h* ranged from 0.44 for B73 × B97 to 0.81 for B73 × Ki3 and B73 × NC350, with the mean of all families = 0.64. Heritability for the combined ten NAM families and repeated parental check lines (ˆ2 *C h* = 0.87) was 23% higher than the average of individuals within-family heritabilities, which represents genetic variation attributable to the diverse parents (Chandler *et al.*, 2013). Repeatability estimate for β-cryptoxanthin (0.89) was higher than that for β-carotene (0.56), reflecting the fact that the environmental influence was larger on the latter trait. The high parent heterosis (HPH) for PVA concentration ranged from −0.36 to 1.00, and the average HPH did not differ ($P = 0.05$) for matings among putative heterotic groups (average HPH = 0.16) and matings of lines within putative heterotic groups (average HPH = 0.06).

Genotype-by-environment interactions for carotenoid concentrations

Genotype-by-environment interaction (GEI) effects can influence the inheritance of carotenoids and their associated QTL. It is therefore expected that effects of genomic regions linked to carotenoid concentration should be estimated for each environment (Zhang *et al.*, 2008). Studies involving many yellow maize tropical inbred lines sampled from four trials conducted in one location and a fifth trial executed in two locations revealed that carotenoid concentrations of lutein, zeaxanthin, β-carotene, β-cryptoxanthin, α-carotene and total PVA contents were not strongly affected by the differences in replications or locations or GEI (Menkir *et al.*, 2008). There was, however, significant genetic-by-location interaction for lutein and PVA content, which was attributed to the magnitude of variation among lines within each location.

In another study, Menkir and Maziya-Dixon (2004) obtained no significant GEI for

β-carotene, among 17 genotypes evaluated in three locations and 2 years. Although Egesel *et al*. (2003a) found that GCA × year interaction for β-carotene was statistically significant, it was of little practical importance (0.75% of the total variation). Suwarno *et al*. (2014) reported that the correlation coefficients among environments were highly significant, indicating a minor role of GEI in the expression of most of the carotenoid component traits. Halilu (2016) also reported that crosses-by-location interaction effects were not significant for PVA components. However, both GCA-by-location and SCA-by-location interaction effects were significant for zeaxanthin; only SCA-by-location interaction effects were significant for lutein. Based on these reports, along with the conclusions by Pfeiffer and McClafferty (2007) and our own findings, it may be concluded that PVA expression is more influenced by genotype and environment than by GEI effects.

A few studies relating kernel colour to PVA and/or β-carotene have been conducted. In the study involving early maturing PVA materials developed at IITA, there was a weak but statistically significant positive correlation between kernel colour and PVA, and also β-carotene, suggesting that to some extent, the degree of the orange colour of kernels could be a quick, though not necessarily the most reliable, strategy to identify inbred lines with high PVA levels. This observation contradicts the findings of Azmach *et al*. (2013), who found no significant correlations between kernel colour and PVA as well as β-carotene contents in the set of intermediate and late-maturing inbred lines studied. The result of the correlation between kernel colour and PVA levels in the two groups of studies suggested that chemical analysis would be the most reliable approach to monitor and improve the levels of the carotenoids, particularly during the early breeding stages.

Population improvement and development of open-pollinated varieties

In 2017, the early and extra-early sub-unit of the IITA Maize programme initiated a study geared towards the development of early (90–95 days to maturity) and extra-early (80–85 days to maturity) stress-tolerant (drought, low-N-tolerant and *Striga*-resistant), along with high PVA, varieties for West and Central Africa (WCA). Towards achieving this objective, crosses were made between four *Striga*-resistant yellow/orange varieties and two sources of high PVA inbred lines to produce eight top-cross hybrids. The varieties were 2004 TZE E-Y STR C_4, TZEE-Y STR QPM, 2004 TZE-Y Pop DT STR C_4 and TZE-Y Pop DT STR QPM, while the inbred lines were Syn –Y-STR-34-1-1-1-1-2-1-B-B-B-B-B/NC354/SYN-Y-STR-34-1-1-1 (OR1) and KU1409/DES/1409 (OR2) obtained from the IITA Maize Improvement Programme. The objective was to introgress genes for high β-carotene into each of the four varieties. Subsequently, the top-cross hybrids were each backcrossed to the respective populations to recover earliness, which resulted in BC_1F_1 progenies. The kernels of the BC_1F_1 of each material with deep orange colour and/or appropriate endosperm modification under the light box in the case of the QPM materials were selected and self-pollinated for two cycles for advancement to the BC_1F_3 stage. Furthermore, BC_1F_3 lines with the deep orange colour were selected and recombined to form the extra-early PVA varieties, 2009 TZEE-OR1 STR, 2009 TZEE-OR2 STR, 2009 TZEE-OR1 STR QPM and 2009 TZEE-OR2 STR QPM; and the early PVA varieties, 2009 TZE-OR1 STR, 2009 TZE-OR2 STR, 2009 TZE-OR1 STR QPM and 2009 TZE-OR2 STR QPM. Because of fund limitations, the PVA varieties have not been screened to determine the levels of β-carotene. However, the varieties have been evaluated in the regional uniform variety trial (RUVT) under multiple contrasting environments since 2010 and several of them have shown outstanding performance. For example, the extra-early varieties 2009 TZEE-OR1 STR and 2009 TZEE-OR2 STR QPM, respectively, out-yielded the extra-early OPV check 2000 SYN EE-W STR by 24 and 9% in the RUVTs conducted across seven environments in WCA in 2015 (Table 17.1). Similarly, two early varieties, 2009 TZE-OR1 DT STR and 2009 TZE-OR1 DT STR QPM, respectively, out-yielded the commercial early OPV check, TZE Comp 3 DT C_2 F_2 by 12 and 11%, across eight contrasting environments in WCA, while 2009 TZE-Y Pop DT STR and 2009 TZE-OR_2 DT STR yielded as much as the early OPV check (Table 17.2).

Table 17.1. Grain yield and other agronomic characters of provitamin A extra-early maturing varieties at seven locations[a] in West and Central Africa.

Variety	Grain yield (kg ha^{-1})	Days to anthesis	Days to silk	Anthesis silking interval	Plant height (cm)	Ear height (cm)	Root lodging (%)	Stalk lodging (%)	Husk cover	Plant aspect	Ear aspect	Ear rot	Ears/plant
2009 TZEE-OR1 STR	4119	53	55	2	166	79	4	5	3	3	4	3	1.1
TZEE-W POP DT C1 STR C5	3999	51	53	2	165	78	4	5	3	3	4	3	0.9
TZEE-W STR 104 BC2	3984	52	54	2	155	77	2	4	3	3	4	3	1.0
2009 TZEE-OR2 STR QPM	3627	53	55	2	160	77	4	5	3	4	5	4	0.9
TZEE-W STR 107 BC2	3596	54	56	2	163	73	4	6	2	3	4	2	1.0
2008 TZEE-Y STR	3380	52	54	2	162	77	4	7	3	4	4	4	0.9
2000 SYN EE-W STR (RE)	3328	51	54	3	163	74	3	7	3	4	5	5	0.9
2009 TZEE-OR1 STR QPM	2501	53	55	3	168	72	7	11	3	5	5	5	1.0
2009 TZEE-OR2 STR	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
GRAND MEAN	**3432**	**53**	**55**	**2**	**163**	**77**	**5**	**7**	**3**	**4**	**4**	**3**	**1.0**
LSD **(α = 0.05)**	**299**	**1**	**1**	**0**	**11**	**5**	**2**	**3**	**1**	**0**	**0**	**1**	**0.1**
CV **(%)**	**16**	**2**	**2**	**37**	**13**	**12**	**34**	**30**	**44**	**18**	**13**	**65**	**24**
Variety	[c]	[c]	[c]	[c]	**ns**[d]	[c]	[c]	[c]	**ns**	[c]	[c]	[c]	**ns**
Environment	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
Variety × Environment	[c]	[c]	[c]	[c]	**ns**	[c]	[c]	[c]	**ns**	[c]	[c]	[c]	[c]

[a]The locations were Ikenne, Bagauda, Mokwa, Zaria, Angaradebou, Bagou, Ina.
[b]Missing data.
[c]Significant at the 1% probability level.
[d]ns = non-significant.

Table 17.2. Means of grain yield (kg/ha) and other agronomic characters of provitamin A early varieties evaluated in 2015 Regional Uniform Variety Trial (RUVT) Early at eight locations[a] in West and Central Africa.

Variety	Grain yield (kg/ha)	Days to anthesis	Days to silk	Anthesis silking interval	Plant height (cm)	Ear height (cm)	Root lodging (%)	Stalk lodging (%)	Husk cover	Plant aspect	Ear aspect	Ear rot	Ears/plant
TZE-W Pop DT C5 STR C5	4556	51	53	2	181	88	2	6	2	3	4	2	1.0
2012 TZE-Y Pop DT C4 STR C5	4459	52	54	2	178	119	3	6	3	3	4	3	1.0
2009 TZE-OR2 DT STR QPM	4021	51	54	2	181	85	3	4	3	3	4	2	1.0
2012 TZE-W Pop DT C4 STR C5	3807	52	54	2	177	86	4	5	3	3	4	3	0.9
2009 TZE-OR1 DT STR	3784	53	55	2	182	86	4	4	3	3	4	2	1.0
2009 TZE-OR1 DT STR QPM	3735	52	54	2	173	85	3	4	3	3	4	2	0.9
2009 TZE-Y Pop DT STR	3666	52	54	2	179	84	3	4	2	4	4	3	0.9
2009 TZE-OR2 DT STR	3659	52	54	2	180	88	4	7	2	3	4	3	1.0
TZE Comp 3 DT C2 F2 (RE)	3370	52	54	2	170	79	5	4	2	4	4	2	0.9
GRAND MEAN	**3874**	**52**	**54**	**2**	**176**	**86**	**3**	**4**	**3**	**3**	**4**	**2**	**0.9**
LSD **(α = 0.05)**	**302**	**1**	**1**	**0**	**6**	**16**	**2**	**2**	**0**	**0**	**0**	**1**	**0.0**
CV **(%)**	**15**	**2**	**3**	**40**	**7**	**36**	**33**	**32**	**26**	**19**	**14**	**68**	**10**
Variety	[c]	[c]	[c]	**ns**	[c]	[c]	[b]	[c]	**ns**	[c]	[c]	[b]	[c]
Environment	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
Variety × Environment	[c]	[c]	[b]	**ns**	**ns**	[b]	[c]	**ns**	**ns**	**ns**	**ns**	[c]	**ns**

[a]Angaradebou, Bagou, Bagauda, Ina, Ikenne, Mokwa, Sekou, Zaria.
[b, c]Significant at the 5% and 1% probability level, respectively.

Furthermore, a programme was initiated in 2011 to extract new generation of extra-early and early maize inbreds from the high PVA normal endosperm varieties, 2009 TZEE-OR$_1$ STR and 2009 TZE-OR$_1$ STR, and the high PVA-QPM varieties, 2009 TZEE-OR$_2$ STR QPM and 2009 TZE - OR$_2$ DT STR QPM. Following the development of the PVA inbred lines, genetic studies were conducted to determine the combining ability of the lines and to identify inbred testers. Several outstanding inbred lines were identified and crossed among themselves to develop biparental crosses, from which inbred lines have been developed through pedigree selection. Additionally, the inbred testers were crossed to the OPVs to develop top-cross hybrids. The top-crosses were outstanding, yielding as highly as the commercial hybrid check, TZEI 124 × TZEI 25 (Table 17.3).

Moreover, a total of 34 PVA extra-early maize hybrids comprising normal endosperm yellow and orange grain colour types, and one commercial hybrid check were evaluated in the IITA Regional Trials under *Striga*-infested, low-N and optimal environments in 2018 (Table 17.4). Grain yield ranged from 1857 kg ha^{-1} for 2009 TZEE-OR$_1$ STR × TZEEI 82 to 3554 kg ha^{-1} for TZEEIOR 197 × TZEEIOR 205 across stress environments, and varied from 3522 kg ha^{-1} for the commercial hybrid check, TZEE-Y Pop STR C5 × TZEEI 58 to 5655 kg ha^{-1} for TZEEIOR 197 × TZEEIOR 205 across non-stress environments. Of interest was the performance of top-cross hybrids involving two extra-early PVA varieties (2009 TZEE-OR$_1$ STR and 2009 TZEE-OR$_2$ STR) and other inbred lines. Three top-cross hybrids, 2009 TZEE-OR$_1$ STR × TZdEEI 7, 2009 TZEE-OR$_1$ STR × TZEEI 67 and 2009 TZEE-OR$_1$ STR × TZdEEI 12, ranked among the top five hybrids in grain yield across stress environments, out-yielding the commercial PVA hybrid check, TZEE-Y Pop STR C$_5$ × TZEEI 58 by 31, 28 and 28%, respectively. However, the top-cross hybrids involving the PVA variety 2009 TZEE-OR$_2$ STR yielded as high as the commercial hybrid check; none of the top-cross hybrids out-performed the commercial hybrid check across stress environments. The outstanding hybrids will be further tested for consistency of performance and commercialized to contribute to food and nutrition security in the subregion.

Molecular Approaches to Provitamin A Enhancement in Maize

Maize kernel colour and MAS have been employed in breeding PVA-rich maize. The discovery of increased nutrition in yellow maize grain led to selection of pigmented grain as a desirable quality trait (Sagare *et al.*, 2015a). It has been reported that high levels of total carotenoids and slightly higher PVA content could be achieved when visual score for kernel colour was used in breeding for PVA-rich maize (Chandler *et al.*, 2013; Menkir *et al.*, 2017; Venado *et al.*, 2017). However, in breeding for increased levels of PVA and total carotenoids in maize kernels, Safawo *et al.* (2010) advocated other more efficient means of quantifying β-carotene in maize grains instead of kernel colour, e.g. expensive HPLC. Muthusamy *et al.* (2015) argued that visual selection based on kernel colour would be inappropriate for selecting PVA-rich genotypes, although improvement in the levels of non-PVA carotenoids was possible using kernel colour. They concluded that molecular breeding would be more appropriate for improving the levels of carotenoids in maize. Lycopene epsilon cyclase (LycE), beta-carotene hydroxylase 1 (crtRB 1) and phytoene synthase (PSY) are functional DNA markers that play an important role in the accumulation of PVA in maize (Harjes *et al.*, 2008; Yan *et al.*, 2010; Fu *et al.*, 2013; Sagare *et al.*, 2015a,b), and β-carotene hydroxylase 1 (crtRB 1 3'TE) has been identified as a favourable allele for effecting a two- to ten-fold increase in kernel β-carotene concentration in maize (Babu *et al.*, 2013; Sagare *et al.*, 2015b). Recently, DNA markers, such as lycopene beta cyclase (lcyB) and Zep-SNP (801), have been reported to contribute to accumulation of PVA levels in maize kernels (Venado *et al.*, 2017; Gebremeskel *et al.*, 2018) and could be employed when breeding for improved levels of PVA in maize kernels.

Information on the genetic diversity and relationship of inbred lines is useful for choosing parents and predicting heterosis (Konstantinov and Mladenović-Drinić, 2007). The use of the conventional methods of maize breeding to obtain information on genetic diversity implies the generation of a large number of crosses among lines and their evaluation in field trials. These extensive field studies are expensive and time consuming. However, assigning inbred lines to

Table 17.3. Grain yield and other agronomic traits of early multiple stress tolerant provitamin A hybrids evaluated under across 8 stress and 12 non-stress environments in Nigeria during the 2017 and 2018 growing and dry seasons.

		Grain yield, (kg ha^{-1})		Days to anthesis		Days to silk		Anthesis silking interval		Plant height (cm)		Ear height (cm)		Husk cover		Plant aspect		Ear aspect		Ears/plant		Stay green xtics (10 WAP)	*Striga* damage (10 WAP)	Emerged *Striga* plants (10 WAP)
Entry	Variety	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR			
1	TZEIOR 58 × TZEIOR 108	3682	5399	54	52	55	53	1	0	155	178	71	82	4	3	5	4	4	3	0.9	1.0	3	5	25
4	TZEIOR 57 × TZEIOR 127	3442	5484	53	52	54	53	1	1	161	178	68	77	4	3	5	4	4	3	0.8	0.9	3	4	43
5	TZEI 124 × TZEI 25 (RE)	3176	5765	54	53	56	54	2	1	152	176	66	78	4	4	5	4	4	4	0.8	0.9	3	5	15
3	TZEIOR 108 × 2009 TZE OR2 DT STR	3131	5065	55	53	57	54	2	1	165	184	80	91	4	3	5	4	5	4	0.8	0.9	3	5	19
2	TZEIOR 108 × 2009 TZE OR1 DT STR	3056	5369	54	54	56	54	2	1	166	182	81	87	4	3	5	4	5	4	0.8	0.9	3	4	27
GRAND MEAN		**3297**	**5416**	**54**	**53**	**56**	**54**	**2**	**1**	**160**	**180**	**73**	**83**	**4**	**3**	**5**	**4**	**4**	**4**	**0.8**	**0.9**	**3**	**4**	**26**
LSD **(α = 0.05)**		**545**	**382**	**1**	**0**	**1**	**1**	**1**	**0**	**8**	**5**	**6**	**4**	**0**	**0**	**0**	**0**	**0**	**0**	**0.1**	**0.1**	**0**	**1**	**14**
CV **(%)**		**29**	**15**	**3**	**2**	**3**	**2**	**69**	**81**	**9**	**6**	**13**	**11**	**20**	**18**	**15**	**14**	**18**	**21**	**15**	**13**	**19**	**16**	**19**
P **for Genotype**		**ns**[a]	[b]	[c]	[c]	[c]	[c]	**ns**	[b]	[c]	[b]	[c]	[c]	[b]	[c]	**ns**	**ns**	[b]	[c]	[c]	**ns**	**ns**	**ns**	[c]
P **for Env**		[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
P **for Genotype × Env**		**ns**	[c]	**ns**	[c]	**ns**	[c]	**ns**	**ns**	**ns**	**ns**	**ns**	**ns**	[b]	[c]	**ns**	[c]	**ns**	[b]	**ns**	**ns**	**ns**	**ns**	**ns**

[a]ns = non-significant.

[b, c]Significant at the 5% and 1% probability level, respectively.

Table 17.4. Grain yield and other agronomic traits of extra-early yellow and orange multiple stress tolerant maize hybrids evaluated across stress (*Striga*-infested and low-N) and non-stress environments in Nigeria, 2018.

Entry	Variety	Grain yield, (kg ha^{-1})		Days to silk		Anthesis silking interval		Plant height (cm)		Husk cover		Plant aspect		Ear aspect		Ears/plant		Stay green xtics (10 WAP[b])	*Striga* damage (10 WAP)	Emerged *Striga* plants (10 WAP)
		STR[a]	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR			
29	TZEEIOR 197 × TZEEIOR 205	3554	5655	56	54	3	2	169	188	3	4	4	4	4	4	0.9	0.9	3	5	56
30	2009 TZEE-OR1 STR × TZdEEI 7	2781	4923	54	52	3	1	161	175	4	4	4	5	5	4	0.8	0.9	3	6	106
10	TZEEIOR 11 × TZdEEI 12	2779	4629	53	52	1	0	151	173	4	4	4	4	4	5	0.8	0.9	4	5	61
13	2009 TZEE-OR1 STR × TZEEI 67	2723	4467	54	54	1	1	165	180	4	4	4	5	4	4	0.8	0.9	3	5	98
15	2009 TZEE-OR1 STR × TZdEEI 12	2718	4781	54	52	2	1	158	179	4	4	5	4	5	4	0.9	0.9	3	5	65
12	TZEEIOR 125 × TZdEEI 7	2632	5302	55	53	1	1	142	177	4	4	4	4	5	4	0.8	1.0	3	5	98
9	TZEEI 81 × TZdEEI 12	2627	4848	53	52	2	2	149	173	4	4	5	5	5	4	0.8	0.9	3	5	67
11	TZEEIOR 30 × TZEEI 79	2566	4558	53	52	1	1	163	175	4	4	4	5	4	4	0.9	1.0	3	5	59
3	(TZEEI 95 × TZEEI 79) × TZEEI 81	2530	4771	52	51	2	1	151	172	4	4	5	5	4	4	0.9	0.9	3	5	49
32	2009 TZEE-OR2 STR × TZdEEI 7	2509	5350	53	53	2	1	148	172	4	4	4	4	5	4	0.8	0.9	3	6	91
6	TZEEI 65 × TZdEEI 7	2506	4459	52	51	2	1	142	174	4	4	4	5	5	4	0.9	0.9	3	6	79
18	(TZdEEI 7 × TZdEEI 12) × TZEEI 81	2484	5158	54	52	3	1	150	178	4	4	4	5	4	4	0.7	0.9	3	5	65

Continued

Table 17.4. Continued.

Entry	Variety	Grain yield, (kg ha^{-1})		Days to silk		Anthesis silking interval		Plant height (cm)		Husk cover		Plant aspect		Ear aspect		Ears/plant		Stay green xtics (10 WAP[b])	*Striga* damage (10 WAP)	Emerged *Striga* plants (10 WAP)
		STR[a]	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR	STR	NSTR			
26	TZEEI 81 × TZdEEI 7	2452	4882	53	51	2	1	155	176	4	4	5	5	4	4	0.8	0.9	3	5	74
1	TZdEEI 7 × TZEEI 58	2440	4533	52	51	1	1	149	183	4	4	5	5	5	4	0.7	0.9	3	5	84
22	TZEE-Y Pop STR 106 × TZEEI 79	2413	4124	54	52	2	1	172	187	4	4	5	5	5	4	0.8	0.9	4	5	49
5	TZEEI 87 × TZdEEI 7	2407	5295	53	51	2	1	148	177	3	3	4	4	4	4	0.9	1.0	2	5	74
8	TZdEEI 1 × TZdEEI 12	2406	4135	54	52	2	1	158	181	4	4	4	5	5	5	0.8	1.0	4	5	33
25	TZEEI 66 × TZdEEI 12	2395	4127	54	52	2	1	167	179	4	4	5	5	5	5	0.8	0.9	3	5	51
27	TZEEIOR 30 × TZEEIOR 142	2393	5205	56	55	2	2	161	189	4	4	5	4	5	4	0.9	0.9	3	5	48
14	TZEE-Y Pop STR BC2 × TZdEEI 7	2380	4659	52	52	1	1	157	172	4	4	5	5	5	4	0.8	0.9	3	5	78
33	2009 TZEE-OR2 STR × TZEEI 58	2376	4251	53	52	1	1	173	187	4	4	4	4	4	5	0.7	0.8	3	6	57
21	(TZdEEI 12 × TZdEEI 13) × TZEEI 81	2322	4775	54	53	2	1	162	173	4	4	5	5	5	4	0.8	0.9	4	6	83
7	TZEEI 89 × TZdEEI 12	2295	3773	52	50	2	1	150	163	4	4	5	5	5	5	0.9	0.9	3	6	37
23	TZEE-Y Pop STR 106 × TZEEI 63	2278	3847	53	51	1	1	148	172	4	4	4	5	5	5	0.7	0.9	3	6	76
24	(TZdEEI 7 × TZdEEI 12) × TZEEI 63	2231	3812	52	51	1	1	155	169	4	4	5	5	5	5	0.8	0.9	3	6	53

17	TZEE-Y POP STR 106 × TZEEI 82	2200	4409	53	50	1	0	167	178	4	4	5	5	5	4	0.8	1.0	3	5	79
4	TZEEI 59 × TZdEEI 7	2157	4200	53	52	2	1	144	164	4	4	5	5	5	5	0.8	0.9	3	6	55
34	TZEE-Y Pop STR C5 × TZEEI 58 (RE)	2122	3522	53	52	2	1	167	183	4	4	5	5	5	5	0.7	0.8	4	6	61
19	(TZdEEI 7 × TZdEEI 12) × TZEEI 58	2119	4299	52	50	2	1	163	180	4	4	4	5	5	5	0.8	0.9	4	5	55
16	TZEE-Y POP STR 106 × TZEEI 81	2073	4546	55	53	3	1	156	181	4	4	5	5	5	4	0.7	0.9	3	6	42
20	(TZdEEI 7 × TZdEEI 12) × TZdEEI 9	2022	4486	53	52	1	1	145	175	4	4	4	4	5	5	0.8	0.9	3	5	90
35	Local Check	1996	3723	54	52	2	1	164	179	4	4	5	5	5	5	0.8	0.9	3	6	88
2	TZdEEI 12 × TZdEEI 58	1992	3791	52	50	1	0	143	166	4	4	5	5	5	5	0.9	0.9	4	5	52
28	TZEEIOR 41 × TZEEIOR 97	1866	4371	56	54	3	2	155	184	5	4	6	5	6	4	0.6	0.8	4	5	63
31	2009 TZEE-OR1 STR × TZEEI 82	1857	4648	53	53	1	1	157	184	4	5	5	5	5	4	0.7	0.9	3	6	93
GRAND MEAN		**2389**	**4523**	**53**	**52**	**2**	**1**	**156**	**177**	**4**	**4**	**5**	**5**	**5**	**4**	**0.8**	**1**	**3**	**5**	**68**
LSD **(α = 0.05)**		**473**	**452**	**1**	**1**	**1**	**1**	**15**	**10**	**0**	**0**	**1**	**0**	**1**	**0**	**0.1**	**0.1**	**1**	**1**	**32**
CV **(%)**		**21**	**15**	**3**	**3**	**72**	**103**	**10**	**9**	**13**	**15**	**12**	**10**	**14**	**16**	**14**	**12**	**16**	**13**	**29**
P **for Genotype**		[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	**ns**	[d]
P **for Env**		[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]		
P **for Genotype × Env**		[c]	[c]	**ns**[e]	[d]	**ns**	[c]	**ns**	**ns**	**ns**	[c]	**ns**	[c]	**ns**	[c]	**ns**	**ns**	[c]		

[a]STR = Stress; NSTR = non-stress.
[b]WAP = weeks after planting.
[c, d]Significant at the 1% and 5% probability level, respectively.
[e]ns = non-significant.

heterotic groups based on molecular markers allows the use of more lines and accelerates the hybrid breeding process (Sserumaga *et al*., 2014).

Various types of molecular markers, including restriction fragment-length polymorphisms (RFLPs), amplified fragment-length polymorphisms (AFLPs), simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) can be used to determine the genetic diversity in a set of inbred lines. The SNP DNA markers have emerged as a powerful tool and are widely used because of the low cost per data point, greater abundance, greater stability and lower mutation rate (Semagn *et al*., 2012, 2014). Progress has been made in marker technologies, from assays that measure the size of the DNA fragment to hybridization-based assays with high multiplexing levels. Diversity Array Technology (DArT) markers, developed relatively recently, have overcome the difficulties encountered with gel-based markers (Sansaloni *et al*., 2010) by utilizing hybridization-based methods (Cruz *et al*., 2013). The DArT markers have advantages over other marker systems by offering an inexpensive and high-throughput genotyping technique and wider genome coverage (Sansaloni *et al*., 2010; Sanchez-Sevilla *et al*., 2015). In addition, they allow rapid germplasm characterization; it is independent of sequencing data and can detect single-base changes (point mutations) and small insertions and deletions (Cruz *et al*., 2013; Sanchez-Sevilla *et al*., 2015). Information on the genetic diversity and distance among breeding lines and the correlation between genetic distance and hybrid performance is crucial for determining breeding strategies, classifying parental lines into heterotic groups, defining heterotic patterns and predicting future hybrid performance (Acquaah, 2012). Assigning lines to heterotic groups prevents the development and evaluation of crosses that should be discarded, allowing maximum heterosis to be exploited by crossing inbred lines belonging to different heterotic groups. Several studies have indicated that crosses among inbred lines derived from unrelated heterotic groups have better grain yield performance than those crosses derived from lines belonging to the same group (Moll *et al*., 1965; Melchinger, 1999; Badu-Apraku *et al*., 2016a). The development of successful maize hybrids, therefore, requires identification of heterotic groups. The use of DArTseq markers has been found to be more efficient than the heterotic grouping based on general combining ability of multiple traits (HGCAMT) grouping method in identifying heterotic groups. Obeng-Bio (2019) used DArTseq markers to identify three heterotic groups for selecting early PVA-QPM maize inbred lines; and TZEIORQ 29 was found to be the best tester both as male and female, whereas TZEIORQ 24 was the best as a male tester. The hybrid TZEIORQ 59 × TZEIORQ 11 was identified as the single-cross tester across research environments.

Konate *et al*. (2017) studied the genetic diversity and the population structure of 110 early-maturing PVA maize inbred lines from the IITA maize improvement programme. The inbred lines were evaluated under drought, *Striga*-infested and optimal conditions in 2015 and 2016 in Nigeria. Significant differences were observed among the lines under different research conditions, indicating that the lines were genetically distinct. The genetic distance between the early PVA lines ranged from 0.03 to 0.45 (Table 17.5). The dendrogram obtained with DArT marker data placed the inbred lines into five heterotic groups (Fig. 17.1). The background information presented in the sections, 'Population improvement and development of OPVs' and 'Development of varieties, inbreds and hybrids with enhanced provitamin A content and tolerance to multiple stresses', on the population from which the PVA early-maturing inbred lines were developed provides a better understanding of the groupings. To a large extent, the high genetic diversity observed among the early-maturing PVA maize inbred lines could be attributed to the broad genetic base of 2009 TZE-OR1 STR. The average genetic distance between individuals from cluster 2 and cluster 1 was higher with values of 0.33 and 0.25, respectively, indicating greater genetic diversity within the groups, with clusters 3, 4 and 5 displaying the lowest values of 0.10, 0.16 and 0.14, respectively (Fig. 17.2).

Even though pedigree information is a useful guide, selection and genetic drift during inbreeding may cause divergence between pedigree and genetic constitution (Liu *et al*., 2003). Sanchez-Sevilla *et al*. (2015) reported that DArT analysis reflected parental relationship between lines. The phylogenetic tree of the 110 inbred lines generated using unweighted

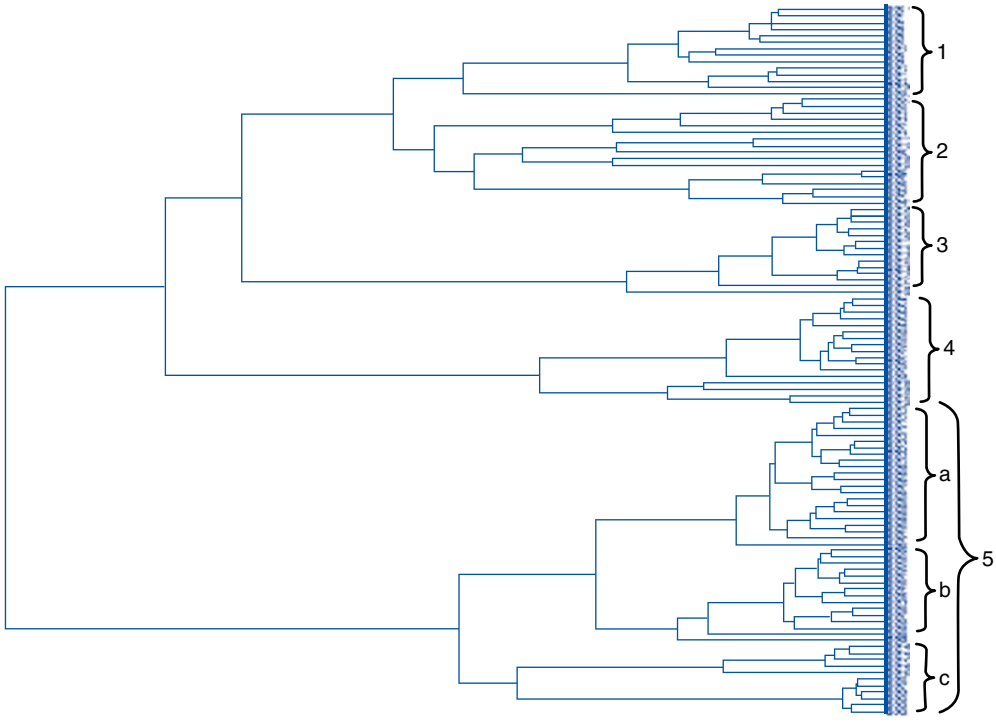


Fig. 17.1. Dendrogram of 110 early provitamin A maize inbred lines based on Diversity Array Technology (DArT) marker data.

neighbour-joining method based on genetic dissimilarity divided the lines into five main clusters (Konate *et al.*, 2017). The lines in cluster 1 were highly diverse, whereas those in each of clusters 2 and 3 were quite similar. Some of the closely related lines were grouped in the same cluster or sub-cluster (Group 3, 5a and 5b), confirming the presence of a relationship between the groupings based on the pedigree and DArT markers in the study (Fig. 17.2). Nevertheless, there were some inconsistencies; for instance, TZEIOR109 and TZEIOR38, TZEIOR92 and TZEIOR110, TZEIOR128 and TZEIOR155 were clustered together despite the fact that they were not closely related by pedigree. Similar findings were reported by Semagn *et al.* (2012). The information generated in this study on the inbred lines would facilitate their further use in the IITA maize breeding programme. There was close correspondence between the cluster analysis and the pedigree information of the inbred lines, indicating that pedigree information can be used effectively for the characterization of early-maturing PVA inbred lines.

In conclusion, the unweighted pair group method with arithmetic mean (UPGMA) clustering method grouped the early-maturing inbred lines into five clusters, which reflected parental relationship between inbred lines despite some inconsistencies. Genetic distance estimates based on DArT markers showed the presence of genetic variation among the early-maturing PVA inbreds that could be useful for hybrid production, population improvement and eventually development of new lines. The information obtained from this study would facilitate better understanding of the genetic relationship among the IITA early-maturity PVA lines.

Agronomic Performance of Provitamin A Maize in Sub-Saharan Africa

Unlike phenotypic marker traits used as descriptors of crop plants because of their consistency of appearance on the plant, quantitative traits

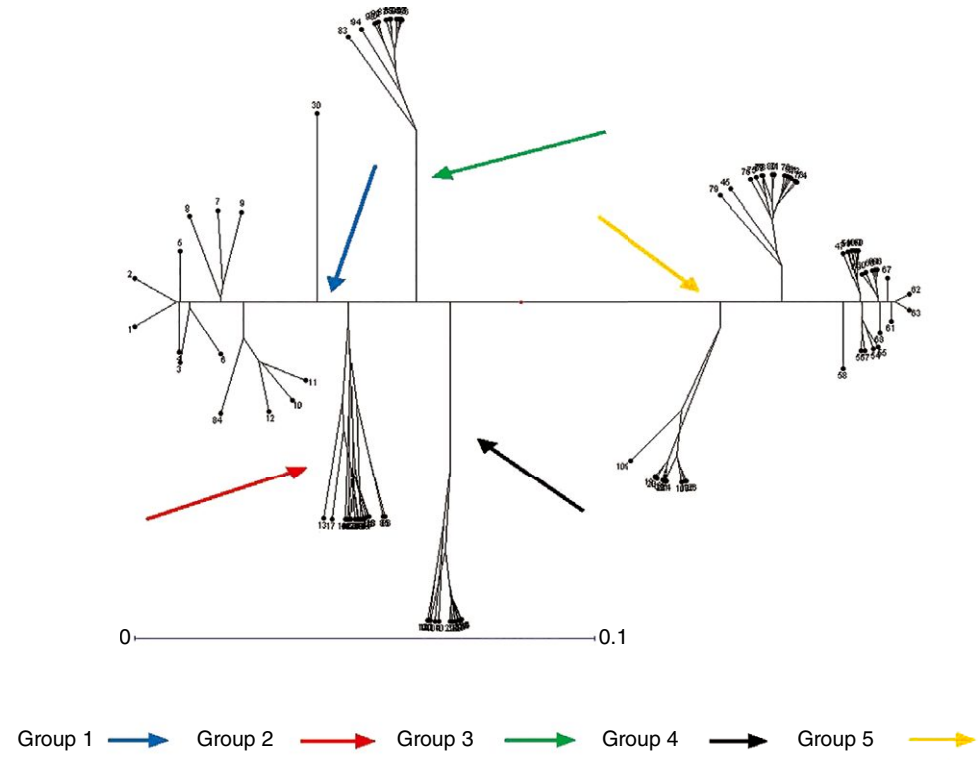


Fig. 17.2. Phylogenetic tree using unweighted neighbour-joining method based on genetic dissimilarity among the 110 early-maturing provitamin A inbred lines.

Table 17.5. Genetic distance (GD) among early provitamin A (PVA) lines represented by the highest and lowest ten divergence combinations.

Ten combinations with low GD	GD	Ten combinations with high GD	GD
TZEIOR51xTZEIOR52	0.03	TZEIOR9xTZEIOR42	0.44
TZEIOR28 xTZEIOR29	0.03	TZEIOR4xTZEIOR56	0.44
TZEIOR90xTZEIOR91	0.04	TZEIOR13xTZEIOR56	0.44
TZEIOR74xTZEIOR75	0.04	TZEIOR13xTZEIOR60	0.44
TZEIOR118xTZEIOR119	0.04	TZEIOR62x TZEIOR13	0.45
TZEIOR120xTZEIOR118	0.04	TZEIOR13xTZEIOR85	0.45
TZEIOR116xTZEIOR117	0.04	TZEIOR13xTZEIOR82	0.45
TZEIOR30xTZEIOR27	0.04	TZEIOR13/TZEIOR87	0.45
TZEIOR25xTZEIOR26	0.04	TZEIOR13/TZEIOR86	0.45
TZEIOR25xTZEIOR29	0.04	TZEIOR13/TZEIOR88	0.45

are seriously affected by environmental factors. For this reason, breeders subject newly developed varieties of crops to extensive agronomic experiments before releasing such varieties for commercial production. For the PVA materials, inbred lines, OPVs and hybrids were evaluated under abiotic stress and non-stressed conditions peculiar to SSA. Although many studies were conducted on OPVs, greater emphasis was on hybrids as the end-products because SSA farmers are gradually moving away from cultivation of OPVs to hybrids. Heterosis, heterotic grouping and combining abilities of inbred lines were studied, and there was an urgent search for

inbred lines and single crosses that could be assessed as testers for the future production of high grain-yielding hybrids with multiple stress resistance/tolerance along with high-quality protein and PVA contents. In addition, data obtained from the agronomic studies were used to determine the interrelations of PVA traits with grain yield and other plant traits in maize.

Development of varieties, inbreds and hybrids with enhanced provitamin A content and tolerance to multiple stresses

As briefly described in the section 'Population improvement and development of OPVs', a programme was initiated in 2007 to develop early and extra-early stress-tolerant (drought, low soil nitrogen-tolerant and *Striga*-resistant), high PVA and quality protein maize (QPM) varieties for SSA. To this end, the extra-early *Striga*-resistant yellow/orange variety, 2004 TZEE-Y STR C4, the extra-early yellow/orange *Striga*-resistant QPM variety, TZEE-Y STR QPM, the early drought- and *Striga*-resistant yellow/orange variety, 2004 TZE-Y Pop DT STR C4, and the drought- and *Striga*-resistant early yellow/orange QPM variety, TZE-Y Pop DT STR QPM, were crossed to two sources of high PVA [Syn –Y-STR-34-1-1-1-1-2-1-B-B-B-B-B/NC354/SYN-Y-STR-34-1-1-1 (OR1) and KU1409/DES/1409 (OR2)] from the IITA Maize Improvement Programme to introgress the genes for high β-carotene into each of the eight varieties. This was followed by a cycle of backcrossing to each recurrent parent to recover earliness. The kernels of the BC_1F_1 of each material with deep orange colour and/or appropriate endosperm modification under the light box in the case of the QPM materials were selected and selfed to advance to the F_2 and subsequently to the F_3 stage. The F_3 lines with the deep orange colour were selected and recombined to form the extra-early PVA varieties, 2009 TZEE-OR1 STR, 2009 TZEE-OR2 STR, 2009 TZEE-OR1 STR QPM and 2009 TZEE-OR2 STR QPM, and the early PVA varieties, 2009 TZE-OR1 STR, 2009 TZE-OR2 STR, 2009 TZE-OR1 STR QPM and 2009 TZE-OR2 STR QPM. Several PVA OPVs developed have been evaluated under both *Striga* infestation and drought since 2010 and have shown outstanding performance under these stresses. Furthermore, a programme was initiated in 2011 to extract a new generation of extra-early and early inbred lines from the high PVA normal endosperm varieties, 2009 TZEE-OR1 STR and 2009 TZE-OR2 STR, and the high PVA-QPM varieties, 2009 TZEE-OR1 STR QPM and 2009 TZE-OR2 DT STR QPM. A total of 155 and 253 inbred lines were extracted from the early and extra-early PVA normal endosperm varieties, 2009 TZE-OR1 STR and 2009 TZEE-OR2 STR, whereas 73 and 76 inbred lines were developed from the PVA-QPM varieties 2009 TZE-OR2 STR QPM and TZEE-OR1 STR QPM, respectively. The inbred lines are presently being used in various genetic studies to (i) determine their combining abilities and heterotic patterns; (ii) classify them into heterotic groups; (iii) identify inbred and single-cross testers; and (iv) determine the performance and stability of the inbreds in hybrid combinations.

It is noteworthy that the original base populations of the TZ (prefix) materials used in the extra-early and early PVA programmes were generated by compositing the available landraces and exotic germplasm from both the landraces and the temperate regions. The landraces collected from Burkina Faso, Nigeria, Mali, Ghana and Mauritania were subjected to extensive testing across WCA for about 2 years by the IITA-WECAMAN (West and Central Africa Collaborative Maize Research Network) programme. The progenies resulting from the composites clearly displayed a mixture of yellow and orange kernel colours, an observation that was casually made at that time, but no attempt was made to separate the colours considered to have no serious genetic implications. Development of inbred lines from the mixture of yellow and orange endosperm colours soon resulted in pure lines with yellow clearly distinct from orange colours, both of which the breeders mistakenly considered to be the same. However, when the HarvestPlus Challenge Programme was initiated, it was found that, on average, the yellow kernels had PVA content of 1.5 μg g^{-1}, whereas the materials with orange colour had 3–8 μg g^{-1}. Attention was then focused on the materials with orange-coloured kernels as a possible source of PVA. Therefore, PVA genes had been in the landraces perhaps in low frequencies. Contrary to the findings of Azmach *et al.* (2013), Tchala (2019)

reported that the orange colour source, STR-34-1-1-1-1-2-1-B*5/NC354/SYN-Y-STR-34-1-1-1-1-2-1-B*5 (OR1), used to convert 2004 TZEE-Y STR C4 to an orange population from which the extra early PVA inbred lines were derived, carried none of the functional alleles of the *crtRB1* gene. He observed that only TZEEIOR 196 and TZdEEI 7 contained the favourable allele at the 3′TE of *lycE* locus while all of the other inbred lines contained the unfavourable allele at 5′TE of *crtRB1* locus. Therefore, the 3′TE functional allele found in TZEEIOR 196 and TZdEEI 7 was not from the donor (i.e. OR1). The presence of this functional allele in these genotypes is a validation of the findings of Yan *et al*. (2010), who reported the 3′TE favourable allele to be present in tropical germplasm at 4.6%.

The WECAMAN programme introgressed resistance/tolerance to drought, *Striga* and low-N into the populations derived from the composites, resulting in stress-tolerant varieties that have been released to farmers in SSA during the past two decades. Much of the breeding efforts for improved PVA presented in this chapter were based on these genetic materials. Indeed, inbred lines with relatively high PVA levels have been developed and used as parent materials for hybrids. Several orange-coloured varieties and hybrids with relatively high PVA levels have been released for commercialization in SSA.

Evaluation of maize inbred lines and hybrids – GCA, SCA, testers and heterotic groupings of early and extra-early PVA inbred lines at IITA

Maize has been successfully introduced into the savannas of SSA principally through extra-early and early maize varieties. Breeding of early and extra-early maize hybrids with high PVA could prevent PVA-deficiency diseases, including night blindness and depressed immune system. In addition, *Striga hermonthica*, drought and low soil nitrogen (low-N) are major stress factors limiting maize production and productivity in savannas of SSA, with a large percentage of the population presently facing PVA deficiency (VAD). However, few commercial extra-early and early-maturing PVA maize hybrids with significant levels of resistance/tolerance to *Striga*, drought and low-N are presently available. Introgression of drought and low-N tolerance as well as *Striga*-resistance genes into high-yielding PVA maize cultivars for *Striga*-endemic and drought-prone areas of the savanna agroecologies of SSA will increase acceptability of PVA cultivars by farmers, promote their adoption, and aid in alleviating VAD and food shortages in the sub-region.

Development and deployment of multiple stress-tolerant maize hybrids with high levels of PVA offer a reliable solution to food insecurity and malnutrition in SSA. Breeding for maize cultivars with elevated levels of PVA carotenoids is a sustainable and effective way to alleviate VAD in SSA. The use of markers for favourable alleles at both *lycE* and *crtRB1* loci in MAS has allowed increases exceeding the breeding target of 15 μg g^{-1} established by the HarvestPlus Challenge Programme for PVA maize hybrids and OPVs. Examples of PVA genotypes having levels of ~17.25 μg g^{-1} (Azmach *et al*., 2013), 15–20 μg g^{-1} (Babu *et al*., 2013) and, more recently, 22.6 μg g^{-1} (Menkir *et al*., 2017) have been reported mostly in maize inbred lines. Despite the excellent progress in breeding for higher levels of PVA, the current released cultivars contain an average of 6–8 μg g^{-1} of PVA (HarvestPlus, 2004; Menkir *et al*., 2017).

A total of 155 and 253 inbred lines extracted from the early and extra-early PVA normal endosperm varieties, 2009 TZE-OR1 STR and 2009 TZEE-OR2 STR and the 73 and 76 inbred lines derived from the PVA-QPM varieties 2009 TZE-OR2 STR QPM and TZEE-OR1 STR QPM, were screened under drought at Ikenne and Bagauda, and under *Striga* infestation at Mokwa. Fifty selected PVA inbreds were analysed for PVA levels in the Food and Nutrition Science Laboratory of IITA-Ibadan. Several outstanding early-maturing PVA, PVA-QPM and extra-early PVA and PVA-QPM inbred lines with elevated levels of PVA were identified for use in the IITA maize breeding programme.

In maize breeding, the per se performance of parental lines is not a reliable indicator of how well or poorly the lines combine and by extension, of the performance of their derived F_1 hybrids. Therefore, combining ability analysis is of major importance in maize breeding, as it helps to determine which maize parental lines

should be selected to improve the local lines and which parent lines should be used in hybrid combinations for high grain yield (Fan *et al.*, 2008). Sprague and Tatum (1942) partitioned the total combining ability of lines into GCA and SCA and defined GCA as the average performance of a line in a series of hybrid combinations and SCA as those instances in which certain hybrid combinations were either better or worse than would be expected based on the average performance of the parent lines studied (Hallauer and Miranda, 1988). Any new germplasm introduced in a breeding programme needs to be tested for GCA and SCA effects. Fan *et al.* (2008) reported that selecting inbred lines with significant and positive GCA effects for grain yield will have a greater chance of producing crosses with higher grain yield.

Mating designs, including diallel (Hayman, 1954; Badu-Apraku *et al.*, 2015b, 2016b), line × tester (Fan *et al.*, 2009; Hosana *et al.*, 2015; Amegbor *et al.*, 2017) and North Carolina Design II (Badu-Apraku *et al.*, 2016a; Annor and Badu-Apraku, 2016) have been extensively used by breeders to determine the relative importance of GCA and SCA for grain yield and its component traits. According to Hallauer and Miranda (1988), GCA variance is related to additive genetic effects, whereas SCA refers to dominance and epistatic gene effects. Additive gene action in reference to a single locus implies lack of dominance, whereas in reference to two or more loci, it refers to the lack of epistasis (Holland, 2001). Fan *et al.* (2008) have used the GCA and SCA mean squares (2GCA MS/(2GCA MS + SCA MS)) ratio to determine whether additive or non-additive gene effects were more important for controlling the inheritance of a trait. The efficiency of selection mainly depends on the additive genetic variation, environment and its interaction with genotype (Zare *et al.*, 2011). The positive GCA effects can be considered favourable or unfavourable, depending on the trait under consideration. Statistically significant positive GCA effects for grain yield and ears per plant indicate high-yielding genotypes. In contrast, significant negative GCA effects for ASI, ear aspect, stay green characteristic, *Striga* damage and number of emerged *Striga* plants are desirable. Plant breeders and geneticists are more interested in genetic variation in a crop and the heritability of desirable traits. Knowledge of the types of gene action involved in controlling the expression of traits is of prime importance for achieving good progress from selection for stress tolerance.

The diallel mating design developed by Griffing (1956) has been routinely employed in genetic studies by plant breeders to obtain information on GCA and SCA of inbred lines (Zhang and Kang, 2003; Zhang *et al.*, 2005). The estimates of GCA and SCA can provide valuable information about the parents used. According to Sughroue (1995), parents that exhibit high GCA effects could be used as testers in hybrid breeding programmes. Therefore, parental lines with highest GCA effects would be expected to produce superior progeny when crossed. Superior hybrids can also be identified by comparing the estimated SCA effects and trait means for each combination (Sughroue, 1995). A better understanding of the mode of gene action controlling the inheritance of important traits is therefore invaluable in hybrid development programmes. The additive gene effects are the predictable portion of the genetic effects (Annor and Badu-Apraku, 2016); therefore, the inheritance through additive gene action allows favourable genes to contribute equally to the improvement of the trait of interest. The non-availability of early-maturing PVA testers, and lack of knowledge of the combining ability and heterotic groups, are major reasons for their non-use in hybrid breeding programmes.

The predominance of additive genetic effects and high heritability for grain yield of inbred lines have been reported (Egesel *et al.*, 2003b), but such information for high levels of PVA in hybrids or OPVs is rare. However, Halilu *et al.* (2016) found non-additive genetic effects to be predominant for all carotenoids. This finding is consistent with that of Burt *et al.* (2011), who reported that heterosis for carotenoids exists, although not so common.

Results of chemical analysis of early-maturing orange and yellow inbred lines at IITA showed that the non-PVA carotenoids (lutein and zeaxanthin) were the predominant carotenoids in the IITA panel of PVA inbred lines (Konate *et al.*, 2017). The concentrations of lutein (14.23–20.26 μg g^{-1}) were higher than those for zeaxanthin (5.25–6.28 μg g^{-1}) in the yellow inbred lines. In contrast, zeaxanthin was the most dominant non-PVA carotenoid in the orange endosperm, with concentrations varying

from 15.71 to 36.42 μg g^{-1}. Among the PVA carotenoids, β-carotene concentrations were higher than those for both β-cryptoxanthin and α-carotene in the orange endosperm inbred lines. However, the level was lower than α-carotene in the yellow endosperm. The provitamin A concentrations in the orange endosperm inbred lines ranged from 3.04 μg g^{-1} for TZEIOR 123 to 9.6 μg g^{-1} for TZEIOR 67, and were higher than those in the yellow endosperm (1.29–1.59 μg g^{-1}). The correlation between carotenoids (β-carotene and α-carotene) and grain yield was not significant (Table 17.6). The results of carotenoid analyses in the IITA early-maturing PVA inbred lines have revealed lower concentrations for β-carotene (0.6–0.8 μg g^{-1}) in the yellow endosperm inbred lines than in the orange endosperm maize (Konate *et al*., 2017). This result is consistent with the findings of Harjes *et al*. (2008), who reported that most yellow maize grown and consumed in the world has low β-carotene concentrations (0.5–1.5 μg g^{-1}). Similar findings with PVA concentrations of less than 2 μg g^{-1} in yellow maize have been reported by many authors. The results of Harjes *et al*. (2008) revealed poor correlations between β-carotene and total carotenoids with orange grain colour.

This implied that the orange grain colour was not always an indication of grains with high or low concentrations of β-carotene. The correlation analyses showed no significant associations of grain yield with α-carotene and β-carotene. The implication is that grain yield was not related to the PVA concentrations. This result corroborated the finding of Menkir and Maziya-Dixon (2004), who observed no significant associations of β-carotene content with grain yield and most of the yield components. Therefore, the development of high-yielding genotypes with improved concentrations of PVA should be possible. About 56% of the inbred lines evaluated in the study by Konate *et al*. (2017) had PVA concentrations ranging from 5 to 9.60 μg g^{-1}. Breeding efforts have resulted in the release of hybrids with total carotenoids concentrations of >7 μg g^{-1} (Dhliwayo *et al*., 2014; Suwarno *et al*., 2014). However, the levels of individual carotenoids observed in the inbred lines were higher than those reported by Menkir *et al*. (2014), but less than the 23.98 μg g^{-1} identified recently in the IITA maize improvement programme for the extra-early PVA inbred TZEEIOR 202 (Badu-Apraku, personal communication). Therefore, selection of parental inbreds with high PVA content and desirable agronomic traits may serve as the basis for developing productive hybrids with high concentrations of PVA. However, 56% of the early-maturing inbred lines used by Konate *et al*. (2017) accumulated PVA concentration levels ranging from 5 to 9.60 μg g^{-1}. It is therefore anticipated that it should to be possible to obtain from the crosses among the early-maturing inbreds evaluated in the present study several hybrids with high PVA concentrations.

Furthermore, only a few released PVA maize hybrids have attained the 15 μg g^{-1} level of PVA established by the HarvestPlus programme in 2004 (Menkir *et al*., 2017). Therefore, several studies have been conducted in IITA on the early- and extra-early-maturing PVA maize to examine the PVA levels of inbred lines and hybrids.

Table 17.6. Correlation between carotenoid[a] concentrations and measured traits[a] of 50 early-maturing provitamin A and yellow inbred lines selected based on their performance under managed drought stress at Ikenne in 2014.

Carotenoids[b]	YIELD	ASI[a]	EASP	EPP	PASP	LD	PLHT	EHT
luT	−0.25NS	−0.21NS	0.24NS	−0.05NS	−0.06NS	−0.16NS	−0.34[c]	−0.44[d]
ZX	0.34[c]	−0.32[c]	−0.34[c]	0.37[d]	−0.21NS	−0.11NS	0.25NS	0.05NS
BCRY	0.37[d]	−0.38NS	−0.41[d]	0.23NS	−0.11NS	−0.09NS	−0.19NS	−0.04NS
AC	0.13NS	−0.44[d]	−0.14NS	0.24NS	−0.14NS	−0.19NS	−0.43[c]	−0.30[c]
BC	0.24NS	−0.44[d]	−0.33[c]	0.40[d]	−0.18NS	−0.1NS	−0.26NS	−0.21NS
ProVitA	0.31[c]	−0.47[d]	−0.38[d]	0.38[d]	−0.18NS	−0.12NS	−0.28NS	−0.18NS

[a]ASI = anthesis-silking interval, EASP = ear aspect, EPP = ears per plant, PASP = plant aspect, LD = leaf death, PLTH = plant height, EHT = ear height.
[b]luT = lutein, ZX = zeaxanthin, BCRY = β-cryptozanthin, AC = α-carotene, BC = β-carotene, and ProVitA = provitamin A.
[c, d]Significant at 5% and 1% probability level, respectively. NS = non-significant.

Combining ability and heterotic patterns of the International Institute of Tropical Agriculture's extra-early and early provitamin A inbred lines

Extra-early inbred lines

Several studies have been conducted under multiple environments to evaluate the combining ability and heterosis of extra-early-maturing PVA inbred lines developed at IITA and the performance of hybrids derived from them. The first set of studies involved 190 F_1 hybrids derived from diallel crosses involving 20 extra-early PVA inbreds plus six checks evaluated using a 14 × 14 lattice design with two replications under *Striga* infestation at Mokwa, under drought at Ikenne, and under optimal environments at Mokwa and Ikenne, 2015–2017. The GCA and SCA effects were computed according to Griffing's method 4 (F_1 hybrids only) (Griffing, 1956). Inbred lines were classified into heterotic groups across environments using the HGCAMT method (Badu-Apraku *et al.*, 2013b). Inbred and single-cross testers were identified and GGE biplot analysis was used to determine the yield and stability of hybrids across environments (Yan *et al.*, 2000). The GCA and SCA effects were significant for grain yield and most other traits, indicating that both additive and non-additive gene actions governed the inheritance of measured traits in F_1 hybrids. However, the dominant effect of the GCA over the SCA effects for the traits suggested that additive gene action was more important than the non-additive in the expression of the traits (Fig. 17.3). Inbred lines TZEEIOR 202 and TZEEIOR 205 had PVA levels of

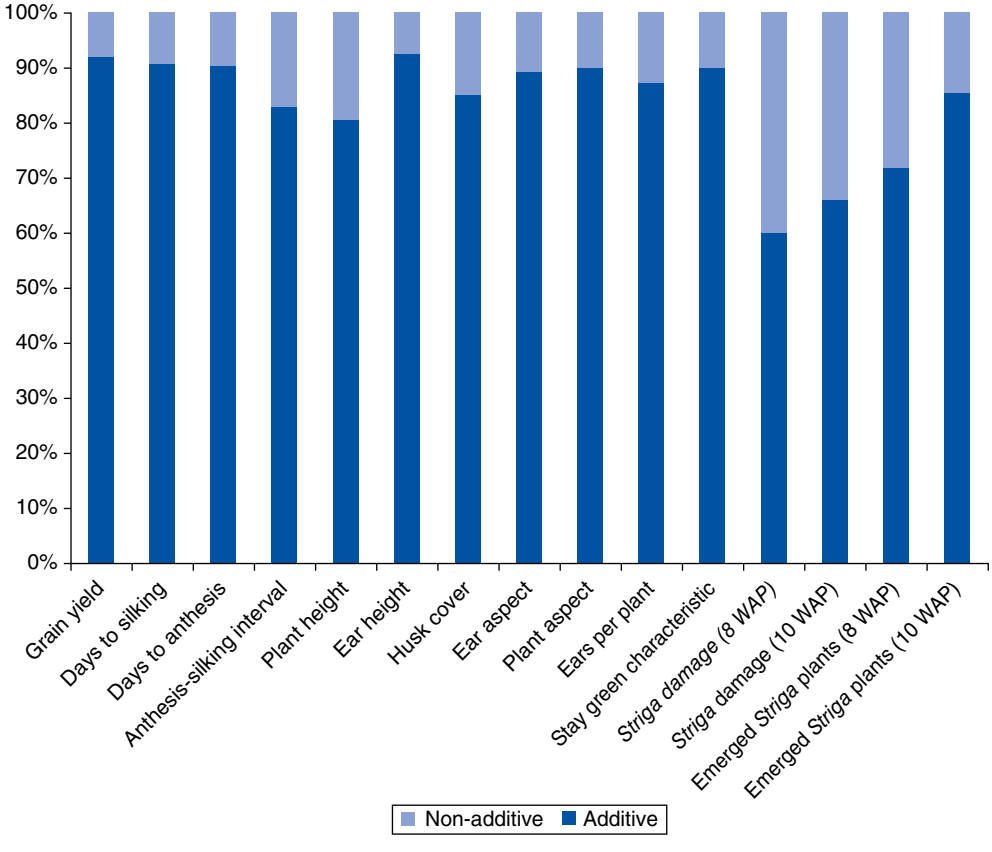


Fig. 17.3. Proportion of additive (lower bar) and non-additive (upper bar) genetic variance for grain yield and other agronomic traits of 20 extra-early provitamin A (PVA) inbred lines involved in diallel crosses evaluated across drought, *Striga*-infested and rainfed environments in Nigeria, 2015–2017.

23.98 and 22.58 μg g^{-1}, respectively (Table 17.7). The HGCAMT method classified the inbred lines into four heterotic groups. The inbreds TZEEIOR 97, TZEEIOR 197 and TZEEIOR 205 were identified as testers for heterotic groups 2, 3 and 4, whereas no inbred satisfied the requirements of a tester for heterotic group 1 (Table 17.8).

Two single-cross testers, TZEEIOR 197 × TZEEIOR 250 and TZEEIOR 205 × TZEEIOR 142, were identified for heterotic groups 3 and 4, respectively. These six hybrids, i.e. TZEEIOR 24 × TZEEIOR 109, TZEEIOR 30 × TZEEIOR 209, TZEEIOR 41 × TZEEIOR 142, TZEEIOR 197 × TZEEIOR 251, TZEEIOR 142 × TZEEIOR 197 and TZEEIOR 30 × TZEEIOR 205, were found to be high yielding and stable across environments (Fig. 17.4) and should be tested extensively and commercialized to contribute to food and nutrition security in SSA. This study resulted in the identification of (i) inbred lines with high levels of PVA that could serve as sources of beneficial alleles for improvement of PVA levels of tropical breeding populations; (ii) inbred and single-cross testers that could be used for classifying PVA inbred lines in SSA; and (iii) high-yielding and stable hybrids that could contribute to both

Table 17.7. Reaction to stresses and provitamin A (PVA) content of extra-early inbred lines used in a diallel study at the International Institute for Tropical Agriculture.

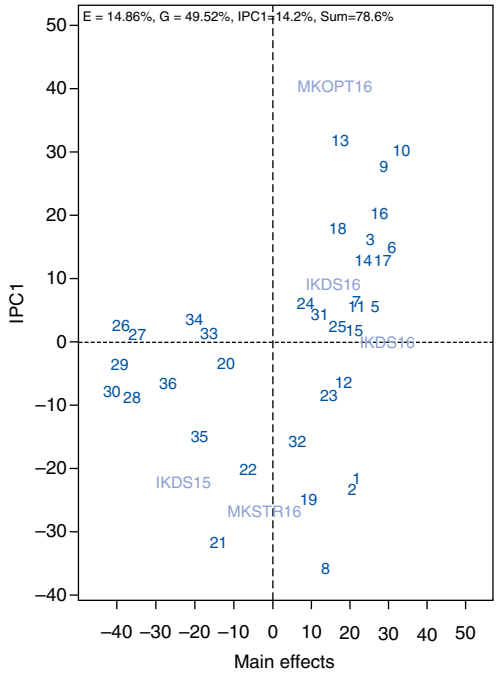
Serial number	Inbred	Reaction to *Striga*[a]	Reaction to drought[a]	Provitamin A content (μg g^{-1})
1	TZEEIOR 22	T	S	9.28
2	TZEEIOR 24	T	S	9.58
3	TZEEIOR 26	S	S	9.74
4	TZEEIOR 27	T	S	7.88
5	TZEEIOR 28	T	T	11.20
6	TZEEIOR 30	T	T	10.19
7	TZEEIOR 41	T	T	11.57
8	TZEEIOR 45	T	T	9.19
9	TZEEIOR 97	T	S	10.44
10	TZEEIOR 109	T	S	10.24
11	TZEEIOR 140	T	S	10.32
12	TZEEIOR 142	T	S	9.86
13	TZEEIOR 197	T	S	8.45
14	TZEEIOR 202	T	T	23.98
15	TZEEIOR 205	T	T	22.58
16	TZEEIOR 209	T	T	9.94
17	TZEEIOR 233	T	S	9.00
18	TZEEIOR 234	T	S	8.33
19	TZEEIOR 250	S	T	8.39
20	TZEEIOR 251	T	T	7.94

[a]T = tolerant/resistant, S = susceptible.

Table 17.8. Heterotic groups of 20 extra-early-maturing provitamin A (PVA) maize inbred lines using the HGCAMT[a] method across eight environments in Nigeria, 2015–2017.

Method	Group 1	Group 2	Group 3	Group 4
HGCAMT	TZEEIOR 22	TZEEIOR 30	TZEEIOR 109	TZEEIOR 140
	TZEEIOR 24	TZEEIOR 97	TZEEIOR 197	TZEEIOR 142
	TZEEIOR 26	TZEEIOR 233	TZEEIOR 209	TZEEIOR 202
	TZEEIOR 27	TZEEIOR 234	TZEEIOR 250	TZEEIOR 205
	TZEEIOR 28		TZEEIOR 251	
	TZEEIOR 41			
	TZEEIOR 45			

[a]Heterotic grouping based on general combining ability of multiple traits.



Code	Hybrids
1	TZEEIOR 30 x TZEEIOR 202
2	TZEEIOR 30 x TZEEIOR 234
3	TZEEIOR 45 x TZEEIOR 205
4	TZEEIOR 109 x TZEEIOR 140
5	TZEEIOR 30 x TZEEIOR 209
6	TZEEIOR 109 x TZEEIOR 197
7	TZEEIOR 41 x TZEEIOR 142
8	TZEEIOR 205 x TZEEIOR 250
9	TZEEIOR 197 x TZEEIOR 205
10	TZEEIOR 109 x TZEEIOR 250
11	TZEEIOR 197 x TZEEIOR 251
12	TZEEIOR 45 x TZEEIOR 142
13	TZEEIOR 27 x TZEEIOR 251
14	TZEEIOR 202 x TZEEIOR 205
15	TZEEIOR 142 x TZEEIOR 197
16	TZEEIOR 30 x TZEEIOR 251
17	TZEEIOR 142 x TZEEIOR 250
18	TZEEIOR 202 x TZEEIOR 250
19	TZEEIOR 97 x TZEEIOR 234
20	TZEEIOR 41 x TZEEIOR 197
21	TZEEIOR 28 x TZEEIOR 109
22	TZEEIOR 41 x TZEEIOR 251
23	TZEEIOR 24 x TZEEIOR 197
24	TZEEIOR 24 x TZEEIOR 109
25	TZEEIOR 30 x TZEEIOR 205
26	TZEEIOR 22 x TZEEIOR 26
27	TZEEIOR 26 x TZEEIOR 41
28	TZEEIOR 26 x TZEEIOR 28
29	TZEEIOR 41 x TZEEIOR 45
30	TZEEIOR 233 x TZEEIOR 234
31	TZdEEI 1 x TZdEEI 12 [Check (CK)]
32	TZdEEI 11 x TZEEI 79 (CK)
33	TZEEI 79 x TZEEI 82 (CK)
34	TZEEI 79 x TZEEI 58 (CK)
35	TZEE-Y Pop STR C5 x TZEEI 82 (CK)
36	TZEE-Y Pop STR C5 x TZEEI 58 (CK)

Fig. 17.4. Mean performance and stability of top 25, worst 5 extra-early provitamin A (PVA) hybrids plus 6 checks in terms of grain yield as measured by principal components across 5 environments in Nigeria, 2015–2017. IKDS15 = Ikenne drought, 2015/16; IKDS16 = Ikenne drought, 2016/17; IKOPT16 = Ikenne optimal, 2016; MKSTR16 = Mokwa Striga, 2016; and MKOPT16 = Mokwa optimal, 2016.

quantity-wise and quality-wise food security in the sub-region.

Maize researchers at IITA have recently made advances in the development of multiple-stress-tolerant maize hybrids with high levels of PVA. At the initial stages of breeding maize for PVA, the HarvestPlus Challenge Programme established 15 μg g^{-1} as the breeding target for PVA maize hybrids and OPVs for commercialization. Till now, only a few released PVA maize hybrids have attained this level of PVA. Recently, the IITA's Early and Extra-early Maize Programme under the leadership of Dr B. Badu-Apraku, in collaboration with other maize scientists and molecular geneticists at IITA and national maize research programmes, has developed extra-early PVA maize inbred lines and hybrids with high levels of PVA. The chemical analysis carried out in the Food and Nutrition Science Laboratory in IITA-Ibadan has shown the following two extra-early PVA maize inbred lines with PVA levels of >22 μg g^{-1}: TZEEIOR 202 (23.98 μg g^{-1}) and TZEEIOR 205 (22.58 μg g^{-1}). Furthermore, crosses involving the high-PVA

maize inbred lines resulted in the development of these high-PVA hybrids: TZEEIOR 197 × TZEEIOR 205 (20.1 µg g^{-1}) and TZEEIOR 202 × TZEEIOR 205 (22.7 µg g^{-1}); these hybrids contained about double the amount of PVA of the commercial PVA hybrid check, TZEE-Y Pop STR C_5 × TZEEI 58 (11.4 µg g^{-1}).

Results of multi-location trials under drought, artificial *Striga* infestation and optimal environments in Nigeria during the period 2015–2017 have shown outstanding agronomic performance of the PVA maize hybrids. The hybrid TZEEIOR 197 × TZEEIOR 205 with PVA level of 20.1 µg g^{-1} yielded 2723 and 4263 kg ha^{-1} across stress (*Striga* and drought) and non-stress environments, respectively (Tables 17.9 and 17.10). In contrast, TZEEIOR 202 × TZEEIOR 205 with PVA level of 22.7 µg g^{-1} yielded 1637 kg ha^{-1} across stress and 4051 kg ha^{-1} across non-stress environments. Hybrid TZEEIOR 197 × TZEEIOR 205 was also identified as the highest yielding and most stable across test environments (Fig. 17.5). The new PVA hybrids out-yielded the commercial PVA top-cross hybrid check, TZEE-Y Pop STR C_5 × TZEEI 58, which yielded 1205 and 2611 kg ha^{-1} across stress and non-stress environments, respectively. These interesting results have offered a great opportunity for breeding and releasing PVA maize hybrids and OPVs with 50% higher levels of PVA than the target of 15 µg g^{-1} established by the HarvestPlus Challenge Programme.

In a second set of studies, 132 extra-early PVA maize hybrids derived from crosses between 33 extra-early PVA inbred lines, along with four inbred testers, were evaluated under *Striga*-infested, drought, low-N and optimal environments in Nigeria, 2015–2016 (Olatise, 2018). Results revealed a preponderance of GCA over SCA for grain yield and other traits under the contrasting environments. Inbred lines TZEEIOR 30, TZEEIOR 41, TZEEIOR 42, TZEEIOR 97, TZEEIOR 109 and TZEEIOR 140 possessed multiple stress-tolerance genes and elevated levels of PVA; these could be used to develop high-PVA, stress-tolerant hybrids. The inbred lines were classified into five groups under multiple stresses and three groups each under optimal and across environments. Inbreds TZEEIOR 197 and TZEEIOR 30 were identified as testers for heterotic groups 1 and 2. Two hybrids, TZEIOR 197 × TZdEEI 12 and TZEIOR 123 × TZdEEI 7, were most stable and high yielding across multiple stress and non-stress environments and are being further tested for commercialization in SSA.

In a third set of field studies, 136 extra-early PVA single-cross hybrids obtained from a 17 × 17 diallel cross, plus four checks, were evaluated by Tchala (2019) in a 10 × 14 lattice design, along with 256 inbred lines in a 16 × 16 lattice, in six environments (two *Striga*-infested, one managed drought and three optimal environments). The objectives were to (i) determine GEI for grain yield and other traits of the 256 extra-early PVA inbred lines; (ii) determine the gene action conditioning grain yield of 17 inbred lines used in the diallel study across contrasting environments; (iii) classify the inbreds used in the diallel into heterotic groups, and identify testers under contrasting environments; and (iv) assess the performance and stability of single-cross hybrids across contrasting environments.

Table 17.9. Results of chemical analysis and grain yield under stress and non-stress environments of extra-early provitamin A (PVA) inbred lines and derived hybrids.

Inbreds	PVA (µg g^{-1})	Reaction to *Striga*	Reaction to drought
TZEEIOR 202	23.98	Tolerant	Tolerant
TZEEIOR 205	22.58	Tolerant	Tolerant
Hybrids	PVA (µg g^{-1})	Yield across drought and *Striga* (kg ha^{-1})	Yield under non-stress (kg ha^{-1})
TZEEIOR 197 × TZEEIOR 205	20.1	2723	4263
TZEEIOR 202 × TZEEIOR 205	22.7	1637	4051
TZEE-Y Pop STR C_5 × TZEEI 58 (Check)	11.4	1205	2611
LSD (α = 0.05)		545	834

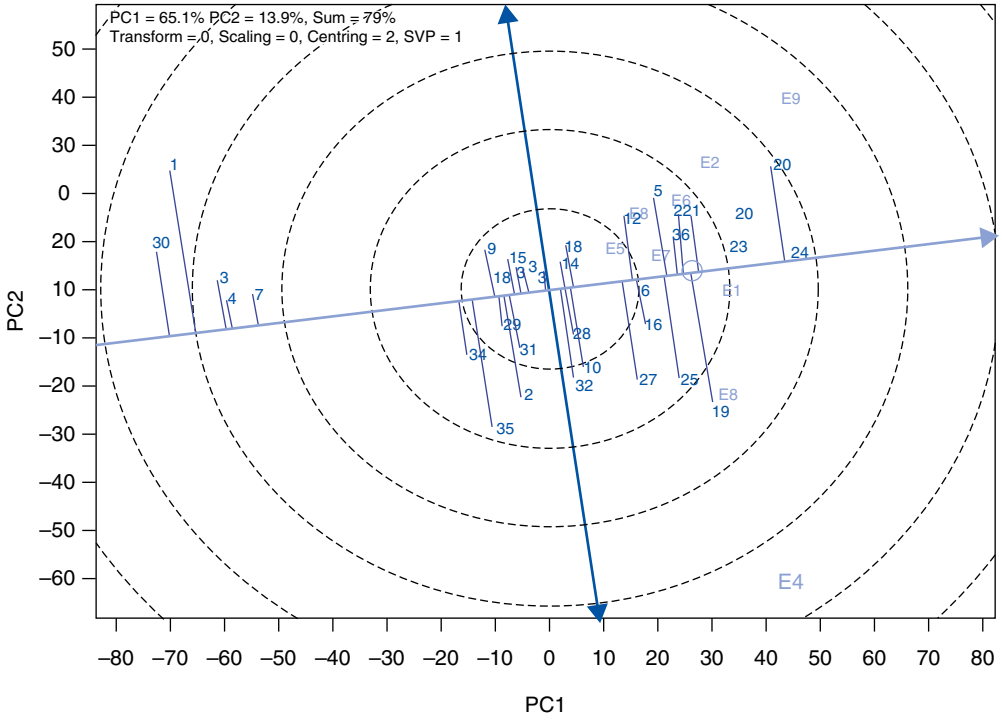
Table 17.10. Performance of selected diallel crosses involving extra-early provitamin A (PVA) inbred lines evaluated under stress (STR) and non-stress (NST) environments in Nigeria, 2015–2017.

	Grain yield (kg ha^{-1})		Days to silk		ASI[a]		Stay green	*Striga* damage	Emerged *Striga* plants
Hybrids	STR	NSR	STR	NST	STR	NST	(10 WAP[b])	(10 WAP)	(10 WAP)
TZEEIOR 109 × TZEEIOR 197	3114	3540	54	52	2	1	4	3	1
TZEEIOR 197 × TZEEIOR 205[c]	2723	4263	55	53	3	1	3	4	1
TZEEIOR 197 × TZEEIOR 251	2559	3238	55	53	2	2	3	4	2
TZEEIOR 140 × TZEEIOR 197	2501	3179	56	56	3	3	4	4	3
TZEEIOR 109 × TZEEIOR 250	2455	3770	54	51	2	1	4	4	1
TZEEIOR 202 × TZEEIOR 205[d]	1637	4051	58	54	4	2	4	5	2
TZEE-Y Pop STR C5 × TZEEI 58 (Check)[e]	1205	2611	55	51	4	1	5	5	1
Mean	**1524**	**2809**	**56**	**53**	**3**	**1**	**5**	**5**	**1**
LSD **(α = 0.05)**	**545**	**834**	**2**	**2**	**2**	**1**	**1**	**1**	**1**

[a]ASI = anthesis-silking interval.
[b]WAP = weeks after planting.
[c]PVA = 20.1 µg g^{-1}.
[d]PVA = 22.7 µg g^{-1}.
[e]PVA = 11.4 µg g^{-1}.

According to the results, TZEEIOR 145 × TZdEEI 7 (4989 kg ha^{-1}), TZEEIOR 12 × TZEEIOR 196 (4835 kg ha^{-1}), TZEEIOR 196 × TZdEEI 7 (4719 kg ha^{-1}), TZEEIOR 196 × TZEEIOR 222 (4197 kg ha^{-1}) and TZEEIOR 161 × TZdEEI 7 (4123 kg ha^{-1}) were the top five outstanding hybrids across test environments. Mean squares attributable to GCA and SCA were highly significant ($P < 0.01$) for grain yield and other traits across environments, except number of ears per plant (EPP) and husk cover. The GCA accounted for about 36 and 43% of the total variation for grain yield across environments and under *Striga*-infested conditions, respectively. The PVA inbred lines TZEEIOR 53 and TZEEIOR 141 of heterotic group II had significant and positive GCA effects for grain yield across environments, thus, making them potential inbred testers. However, the PVA inbred TZEEIOR 141 was identified as tester over TZEEIOR 53, because it had higher positive and significant GCA effect for grain yield and higher yield per se in addition to belonging to a heterotic group. Stability analysis across environments identified TZEEIOR 196 × TZdEEI 7 and TZEEIOR 196 × TZEEIOR 222 as the most stable hybrids, followed by TZEEIOR 145 × TZdEEI 7 and TZEEIOR 12 × TZEEIOR 196.

The fourth study was conducted by Tchala (2019) to (i) examine the breeding values of extra-early maturity PVA inbred lines of the IITA maize programme for resistance to *Striga* and tolerance to drought; (ii) assess the genetic purity and diversity in the selected extra-early maturing PVA inbred lines using SNP-based DArTseq markers; (iii) determine the gene action conditioning *Striga* resistance and tolerance to drought in extra-early maturity PVA inbred lines; (iv) assess the grain yield performance and stability of the hybrids under varying environments; and (v) determine the mode of inheritance of carotenoids in extra-early maturity orange maize inbred lines. In this study, 180 inbred lines, including 152 selected PVA lines, were evaluated under *Striga*-infested, managed drought stress and optimal environments. Based on IITA's base indices for selection, 19% of the total inbreds evaluated, including 21% of the PVA inbreds, combined *Striga* resistance and drought



➢ **TZEEIOR 197 x TZEEIOR 205 – highest yielding and most stable (24).**

Fig. 17.5. GGE biplot for grain yield of selected diallel crosses involving extra-early provitamin A (PVA) inbreds across stress and non-stress environments in Nigeria, 2015–2017.

tolerance. Seventy per cent of the 19%, i.e. 13.3%, of inbreds were also selected based on multivariate best linear unbiased predictors (BLUPs) across all environments. The genetic purity and diversity among the 152 orange inbreds were assessed using 4620 polymorphic SNPs. The results revealed that 92% of the inbreds were pure with heterozygosity <5%, whereas the remaining 8% had heterozygosity ranging from 5.1% to 20.2%. Roger's genetic distance for about 71% of the pairs of lines fell between 0.2001 and 0.2500. Ninety-two per cent of the pairs of lines also showed relative kinship values ranging from 0.300 to 0.500. Model-based population structure analysis and neighbour-joining cluster analysis assigned 71% of the inbreds into four distinct distant groups. Fifteen inbreds selected from among the 152 evaluated, plus TZdEEI 7 and TZdEEI 12, were used to generate 136 diallel single-cross hybrids, which were evaluated together with four experimental hybrid checks under *Striga*-infested, drought-stress and optimal environments at three locations in Nigeria in 2016 and 2017 (11 environments). General and specific combining ability components of the genetic variance were significantly different from zero for grain yield and most of the traits. Additive and non-additive genetic effects were both important, with a predominance of the latter in controlling grain yield and most of the measured traits under *Striga*-infested, drought-stress and across the contrasting environments. However, additive genetic effects were the primary effects modulating the stay-green characteristic and *Striga* resistance-indicator traits, suggesting that selection for these traits could easily be carried out based on predictions of GCA alone. Results also showed that non-additive genetic effects were the primary type of effects for grain yield and most other agronomic traits in all environments, except for the stay-green characteristic under drought stress and *Striga* resistance-indicator traits, which could be improved based solely on GCA effects of

parental lines. Using the IITA base indices, 26% of the hybrids were found to combine *Striga* resistance with drought tolerance. Stability analysis of the top 26 hybrids across test environments based on their genetic values indicated that TZEEIOR 12 × TZEEIOR 196 was the most stable hybrid, combining resistance to *Striga* and tolerance to drought with grain yield (3885 kg ha^{-1} across environments and 5411 kg ha^{-1} under optimal conditions). Superior hybrids identified in this study could successfully be used to improve maize productivity and production, and well-being of less privileged farmers in SSA, especially in *Striga*-endemic environments. Results of this study also indicated that the top five high-yielding and stable hybrids (TZEEIOR 12 × TZEEIOR 196, TZEEIOR 145 × TZdEEI 7, TZEEIOR 222 × TZdEEI 7, TZEEIOR 223 × TZdEEI 7 and TZEEIOR 219 × TZdEEI 7) outyielded the best check (TZEEI 79 × TZEEI 82) by 35–60% across environments. These outstanding hybrids were also among the top-performing hybrids under *Striga* infestation and could, therefore, be tested extensively to confirm their superior performance and commercialized. Different superior hybrids were also identified under different stresses. There is therefore the need for introgression of these new sources of genes for resistance/tolerance to *Striga* and tolerance to drought in tropical PVA breeding populations.

Results of diallel analysis using the Hayman method revealed the presence of more dominant genes than recessive genes in the parents, with the ratio of dominant to recessive genes being greater than 2 for β-carotene (2.36). Also, at the loci exhibiting dominance, recessive alleles were mostly positive for β-cryptoxanthin, lutein and, to some extent, for β-carotene, whereas dominant genes expressed both effects for the rest of the traits. It was concluded that primarily dominance conditioned most of the carotenoid traits in the set of parental inbreds used in the study.

In the fifth study, Oyekale (2019) evaluated 150 TZEEIORQ hybrids, which were developed from hybridizing 30 selected inbreds (in six sets) using the North Carolina Design II, along with six checks at three locations simultaneously, but separately, with parental inbreds tested in 2 years using the alpha lattice design with two replications under *Striga* infestation, low-N (30 kg N per ha) and high-N (90 kg N per ha) conditions. Also, inbreds were screened using a PVA DNA marker and via chemical analyses. Equally, relatedness among inbreds was investigated using DArT-derived SNPs. The objectives of this study were to: (i) identify tropical *Zea* extra-early PVA quality protein maize (TZEEIORQ) inbreds that combine *Striga* resistance and low-N tolerance with elevated levels of PVA, tryptophan and lysine; (ii) assess genetic effects for grain yield, *Striga* resistance, low-N tolerance and other agronomic traits; (iii) categorize the inbreds into heterotic groups and identify testers; (iv) examine the relationships among and between the traits of inbreds and hybrids; and (v) identify high-yielding and stable hybrids across *Striga*-infested, low-N and high-N environments.

There were significant ($P < 0.01$) differences among the inbreds for grain yield and other agronomic traits across environments. Multiple-stress base index was positive for 50% of the inbreds, which were considered resistant/tolerant to both *Striga* infestation and low N. The PVA marker grouped the inbreds into two classes. Levels of PVA, tryptophan and lysine in the inbreds ranged from 2.21–10.95 μg g^{-1}, 0.04–0.08% and 0.19–0.39%, respectively. Inbreds TZEEIORQ 58, TZEEIORQ 55, TZEEIORQ 5, TZEEIORQ 52, TZEEIORQ 57 and TZEEIORQ 62 combined *Striga* resistance and low-N tolerance with elevated levels of PVA, tryptophan and lysine. GCA and SCA mean squares were significant ($P < 0.01$) for grain yield and other agronomic traits across environments. However, GCA was preponderant over SCA for grain yield (86% versus 14%), *Striga* damage (88% versus 12%), number of emerged *Striga* plants (77% versus 23%), stay-green characteristic (75% versus 25%) and other traits, indicating that these traits were largely controlled by additive gene effects. The ratio of GCAf to GCAm was greater than unity for grain yield, *Striga* damage syndrome ratings and stay-green characteristic, suggesting a greater influence of maternal effects on the traits. Ear aspect accounted for about 62% of the variation in grain yield of both inbreds and hybrids under *Striga* infestation. The inbreds were categorized into four heterotic groups using GCA of multiple traits, whereas three heterotic groups were obtained using DArT-derived SNPs. Inbreds

TZEEIORQ 5, TZEEIORQ 53 and TZEEIORQ 61 were identified as good testers. Hybrids TZEEIORQ 49 × TZEEIORQ 75, TZEEIORQ 55 × TZEEIORQ 26 and TZEEIORQ 52 × TZEEIORQ 43 were the most stable and high yielding across environments. These promising hybrids should be tested further and promoted for commercialization in SSA.

Early inbred lines

Based on the fact that VAD is a major health problem in SSA and the premise that the maize plant can accumulate a significant quantity of PVA in the endosperm and has significant genetic variation for the PVA trait, IITA's maize breeding efforts have been focused on improving PVA, along with tolerance/resistance to abiotic and biotic production constraints. The savanna agroecologies, especially the northern Guinea savanna, which is the corn-belt of WCA, is vulnerable to sporadic as well as to terminal drought during the growing season. Therefore, early-maturing maize, into which drought tolerance has been introgressed, best fits this agroecology for optimum grain production. Studies were conducted with the objectives of determining the combining ability of 20 early-maturing PVA maize inbred lines under contrasting environments, classifying the inbred lines into heterotic groups, identifying testers and evaluating the agronomic performance of the inbred lines in hybrid combinations. Field trials were conducted using 190 diallel crosses generated from 20 inbred lines. Six yellow-endosperm hybrid checks were included in the trials. The 196 entries were evaluated under drought, *Striga*-infested, low-N and optimal environments in Nigeria, 2016–2017. Mean squares were significant for GCA and SCA effects for most of the traits across environments, suggesting that both additive and non-additive gene actions governed the inheritance of these traits. The preponderance of GCA effects for most traits over the SCA effects suggested that additive gene effects were more important than the non-additive effects. The PVA inbred lines were classified into three heterotic groups based on HGCAMT method. Inbreds TZEI 25 and TZEIOR 164 were identified as testers for groups 2 and 3, respectively; no inbred tester was identified for group 1. Both inbred testers showed significant positive GCA effects for grain yield across environments. Furthermore, TZEI 25 revealed significant negative GCA effects for *Striga* damage and number of emerged *Striga* plants at 10 weeks after planting (SDR2 and ESP2), whereas significant negative GCA effects were detected for ESP2 of TZEIOR 164. The characteristics of these inbred lines indicated that they could be invaluable sources of beneficial alleles for the development of superior PVA hybrids and populations for the tropics. The GGE biplot identified PVA hybrid TZEIOR 4 × TZEIOR 158 as the highest yielding across environments; it out-yielded the best hybrid check by 35.9% under stress. This hybrid is being further tested for commercialization to improve food security and sustain maize production in SSA.

Several other studies have been conducted at IITA on combining ability, heterosis and genetic diversity of early-maturing PVA maize inbreds under drought-stress and *Striga*-infested environments. Konate *et al*. (2017) studied the reactions of a set of the early-maturing PVA inbreds for tolerance/resistance to drought and *Striga* and determined the PVA contents; examined the combining ability and heterotic groups of selected drought and *Striga*-resistant early PVA inbreds under *Striga*-infested, drought-stress, optimal and across environments; and investigated the genetic diversity and population structure of the PVA inbred lines using the DArT markers. During the 2014 dry season, a set of 155 early-maturing PVA maize inbred lines was screened under managed drought to select 100 promising inbred lines, which were characterized for genetic diversity using SNP markers. The inbreds were also evaluated under drought, *Striga*-infested and optimal environments during the 2015–2016 growing seasons in Nigeria to confirm the consistency of their reactions to the biotic and abiotic stresses. Fifty of the set of 100 inbred lines were analysed for carotenoid content. About 56% of the lines had PVA concentrations ranging from 5 to 9.60 μg g^{-1}. The correlation analyses did not show significant associations of grain yield with α-carotene and β-carotene contents, indicating that high-yielding inbred lines with high PVA level could be selected for hybrid development. Grain yield of the inbred lines across drought and *Striga*-infested environments ranged from 119 to 1971 kg ha^{-1}, with a mean of 893 kg ha^{-1}. Of the 100 early-maturing inbred lines evaluated under

drought, 50 had positive base indices, with 39 of them yielding above the mean grain yield. Under *Striga*-infested environments, the top 20 inbred lines produced above mean grain yield, which varied from 866 to1340 kg ha^{-1}. Cluster analysis of the genetic distance classified the lines into five groups based predominantly on their pedigrees. Results of diallel analysis involving 17 selected inbred lines (136 single-cross hybrids) showed that the hybrids exhibited significant differences for all measured traits under drought, *Striga*-infested, optimal and across stress environments, except for ASI under *Striga*-infested environments. The GCA and SCA effects were significant for grain yield and other traits under drought, *Striga*-infested, optimal and across environments, except for SCA effects for ASI and EPP under drought and *Striga* environments. Significant positive GCA effects were observed for grain yield for TZEIOR 108, TZEI 10 and TZEI 17 across stress and non-stress environments. In addition, significant negative GCA effects were obtained for ear aspect and plant aspect for TZEIOR 108, TZEI 17 and TZEI 10 under stress and non-stress environments. The inbreds TZEIOR 108 and TZEI 10 had significant negative GCA effects for *Striga* damage and number of emerged *Striga* plants. The preponderance of GCA effects over SCA effects for grain yield and most other measured traits under drought, *Striga*-infested, optimal and across environments suggested that additive gene effects were more important in the inheritance of these traits. This implied that good parents could be identified using the measured traits and promising hybrids could be produced based on the prediction from GCA effects. The HSGCA grouping method classified the lines into four main groups. The inbred lines TZEIOR 108 and TZEI 10 were identified as the best testers. The GGE biplot analysis identified the hybrids TZEIOR 57 × TZEIOR 108, TZEIOR 13 × TZEIOR 59, TZEIOR 60 × TZEIOR 108, TZEIOR 127 × TZEI 10, TZEIOR 9 × TZEIOR 56 and TZEIOR 58 × TZEIOR 108 as the highest yielding and most stable across environments. The hybrids TZEIOR 60 × TZEIOR 127 and the commercial hybrid check TZEI 124 × TZEI 25 were high yielding but unstable across environments.

A study was conducted by Obeng-Bio (2019) using the IITA early-maturing PVA-QPM inbred lines to (i) identify drought and low-N-tolerant inbred lines with elevated levels of PVA and quality protein; (ii) assess the extent of genetic diversity and population structure of selected early-maturing drought and low-N-tolerant inbred lines; (iii) determine combining ability of the inbred lines for drought and/or low-N tolerance, and PVA and tryptophan accumulation; (iv) classify the inbred lines into heterotic groups and identify the best inbred and hybrid testers across environments; (v) assess yield and stability of hybrids across stress and non-stress environments; and (vi) validate the presence of PVA functional genes in the set of inbred lines. The secondary traits measured complemented the grain yield of the set of inbred lines in identifying 33 (out of the 70) drought and low-N-tolerant inbred lines for the genetic studies. Moderate levels of PVA were recorded for the inbred lines assayed, indicating the need to introgress the best favourable PVA alleles into the same population from which the inbred lines were extracted. The inbred lines TZEIORQ 55 and TZEIORQ 29 were the best in PVA contents, recording 15.38 and 12.10 μg g^{-1}, respectively, whereas inbred lines combined moderate levels of PVA with drought and low-N tolerance.

Breeders have successfully used mating designs, such as diallel, line × tester and the North Carolina Design II to establish heterotic groups for maize inbreds (Hosana *et al.*, 2015; Badu-Apraku *et al.*, 2016a). However, morphological traits, especially quantitative traits, do not allow detection of differences among closely related genotypes and are strongly influenced by prevailing environmental conditions (Smith and Smith, 1992). In view of this, molecular markers have been used as a more powerful option for classifying inbreds into heterotic groups (Barata and Carena, 2006). This method is extremely helpful in instances of new inbred sets with no pedigree information. Several reports have demonstrated a high correlation between genetic distance and hybrid performance in maize (Lee *et al.*, 1989; Smith *et al.*, 1990; Betrán *et al.*, 2003; Xu *et al.*, 2004; Kiula *et al.*, 2008). Contrary to these reports, some other authors have reported that genetic distance measures are of limited use in predicting hybrid performance, heterosis and SCA of single crosses (Melchinger *et al.*, 1990). The general underlying explanation for this phenomenon is that although molecular markers are highly efficient for mapping the location of genes in the genome, they provide little

information on the physiological functions and interactions of gene products. Heterosis is based on dominance and epistatic interactions of gene products, which DNA markers may not measure accurately. Because of the shortcomings of both morphological traits and molecular markers in classifying inbreds into heterotic groups, any of the approaches could be used in combination to complement the other in a genetic improvement programme.

Different methods of identifying heterotic groups have been proposed. The use of SCA effects of grain yield to classify inbreds into heterotic groups (Fan *et al*., 2004) is recognized as the oldest method and it has been employed in many studies (Menkir *et al*., 2004; Badu-Apraku *et al*., 2015b). However, using SCA effects for grain yield to assign inbreds to different heterotic groups could be biased by genotype × environment interactions, which could be the cause of inconsistent grouping of the inbred lines in different studies (Badu-Apraku *et al*., 2013b). The SCA and GCA effects for grain yield were therefore combined in a method called heterotic group's specific and general combining ability effects of grain yield (HSGCA), which is regarded as a more efficient approach for grouping maize inbreds (Fan *et al*., 2009). Unfortunately, the SCA and HSGCA methods are based only on grain yield, making them less efficient. This is because grain yield has low heritability, especially under stress, as indicated by Bolaños and Edmeades (1993), and hence directly selecting for grain yield alone under drought might delay progress. For these reasons, Badu-Apraku *et al*. (2013b) proposed heterotic grouping based on the HGCAMT method. Only the measured traits with significant GCA effects are employed in this method, and it becomes the method of choice for classifying inbreds into heterotic groups in the NCD II arrangements, where the crosses are restricted to specific sets of inbred lines. Furthermore, it is an appropriate method for grouping inbred lines when the breeding objective is to develop resistance or tolerance to multiple stresses in trials involving the measurement of several traits.

Several authors have assessed the efficiency of the different heterotic grouping methods and the results so far have been contradictory. For example, Fan *et al*. (2009) and Akinwale *et al*. (2014) found the HSGCA method to be the most efficient after comparing the SCA, HSGCA and SNP marker methods. Similarly, Badu-Apraku *et al*. (2015b) identified the HSGCA method as the most efficient, followed by the HGCAMT, SNP markers and the SCA in a study that compared the efficiencies of these four grouping methods. Also, the HSGCA method was more efficient than the SCA method in classifying the extra-early maturing white maize inbred lines into heterotic groups (Amegbor *et al*., 2017). Conversely, Badu-Apraku *et al*. (2013a, 2016b) ranked the HGCAMT method as the most efficient, followed by HSGCA and then the SNP-based method. The contradictory reports emanating from the different studies were principally ascribed to the differences in the genetic materials used and their responses to the prevailing environmental conditions with respect to the SCA, HSCGA and HGCAMT methods, as well as the number of markers used with respect to the SNP-marker-based method. Using DArTseq markers, UPGMA clustering, model-based structure analysis and principal component analysis were employed to assess the genetic diversity among the inbred lines. The results consistently revealed five clusters, which were based largely on the pedigrees of the set of inbred lines, indicating the existence of genetically distinct groups.

Twenty-four PVA-QPM early-maturing inbred lines were selected to generate 96 North Carolina Design II crosses, which were evaluated under drought, low-N and optimal environments in Nigeria, 2015–2018 (Obeng-Bio, 2019). Results revealed that the additive genetic effects were more important than the non-additive effects for grain yield and most other agronomic traits under individual environments and across environments. Maternal effects were not significant for measured traits under drought, low-N, optimal and across environments as well as for the carotenoids and tryptophan contents assayed, except for the stay-green characteristic and ASI across environments. In addition, paternal effects significantly ($P < 0.05$) conditioned the inheritance of prolificacy under optimal conditions.

Maize breeders at IITA use a base index to select genetic materials improved for several traits simultaneously. In one such study, 32 of the 70 early-maturing PVA-QPM inbred lines evaluated were identified as drought tolerant on the basis of the drought base index. These inbreds would serve as an important source of

genes for the development of superior drought-tolerant hybrids (Betrán *et al.*, 2003), synthetics and for the improvement of the early-maturing PVA-QPM population for drought tolerance. Similarly, 37 low-N-tolerant inbred lines were identified, which would be crucial for the exploitation of low-N-tolerance genes to develop superior hybrids and synthetic varieties under low-N conditions (Adofo-Boateng *et al.*, 2015). Moreover, 33 inbred lines were identified as drought- and low-N-tolerant based on the multiple trait base index, suggesting that similar adaptive mechanisms were involved in the tolerance to the two stresses and that selection under drought- could also improve low-N tolerance, as reported by several workers (Kim and Adetimirin, 1997; Bänziger *et al.*, 1999; Badu-Apraku *et al.*, 2012). In other studies, the PVA levels of 18 selected early-maturing PVA-QPM parental lines ranged from 3.47 to 15.38 μg g^{-1}, with a mean of 6.47 μg g^{-1}, indicating the existence of significant variation for the PVA carotenoids in the set of inbred lines used (Weber, 1987; Pfeiffer and McClafferty, 2007; Harjes *et al.*, 2008; Mishra and Singh, 2010). This range of PVA values exceeded the 5.00 to 7.80 μg g^{-1} range reported by Menkir *et al.* (2008) from 15 tropically adapted yellow maize inbred lines but was similar to the 0.06 to 17.25 μg g^{-1} range, with a mean of 5.87 μg g^{-1} reported by Azmach *et al.* (2013) using 130 inbred lines. However, only TZEIORQ 55 recorded a PVA value >15 μg g^{-1}, which is the breeding target established by HarvestPlus (Ortiz-Monasterio *et al.*, 2007; Harjes *et al.*, 2008; HarvestPlus, 2004). Although the results indicated the potential of achieving the established target using this inbred set, there is the need to introgress the best favourable alleles for PVA from the outstanding tropical germplasm sources, such as TZEEIOR 2002 and TZEEIOR 2005, into the tropically adapted inbred lines to facilitate the development of high PVA-QPM hybrids with good adaptation to drought and low-N environments. The highest estimated mean of total carotenoids was 60.22 μg g^{-1}, which was higher than the 42.71 μg g^{-1} reported by Azmach *et al.* (2013) but far below the 100 μg g^{-1} reported by Burt *et al.* (2011). The significance of high total carotenoids is that inbred lines harbouring higher amounts of total carotenoids could be invaluable sources of the PVA carotenoids, especially if the influx of assimilates to the carotenoid biosynthetic pathway favours the accumulation of the PVA carotenoids in the endosperm. Also, the results revealed relatively high levels of lutein and zeaxanthin (synthesized from the PVA carotenoids) at the expense of the PVA carotenoids for most of the inbreds. This result is consistent with the report by Howitt and Pogson, (2006), who identified lutein and zeaxanthin as the most predominant carotenoids in the maize endosperm. This result, however, is inconsistent with the findings of Babu *et al.* (2013), who reported many genotypes having high PVA contents (ranging from 15 to 20 μg g^{-1}) compared with the non-PVA carotenoids when improved PVA inbred lines and populations were studied. Ultimately, new sources of PVA genes would be necessary to improve the existing early-maturing PVA-QPM inbred population to speed up the development of the next generation of high-PVA tropical maize hybrids for commercialization in SSA to combat VAD. The two inbred lines, TZEIORQ 55 and TZEIORQ 29, which possessed high levels of PVA (15.38 and 12.10 μg g^{-1}, respectively) and low-N tolerance (for TZEIORQ 29) could be invaluable sources of PVA genes for the improvement of the early PVA-QPM source inbred population. PVA of the inbreds did not correlate with grain yield, suggesting that the two traits can be improved simultaneously in the set of inbred lines. This result might be the reason why most of the inbreds with combined drought and low-N tolerance generally recorded low levels of PVA. This finding, therefore, suggested that the inbreds were relatively better adapted to tropical environments relative to drought and low-N, but the PVA levels needed improvement. Also, the non-significant correlations observed among PVA and lutein and zeaxanthin indicated that the levels of the PVA carotenoids (β-carotene, β-cryptoxanthin and α-carotene) could be improved without significant loss associated with the synthesis of lutein and zeaxanthin in the PVA biosynthetic pathway.

The GGE-biplot analysis and the drought and low-N multiple trait base index consistently identified TZEIORQ 24 × TZEIORQ 41 as the highest yielding and most stable hybrid across stress and non-stress environments. TZEIORQ 29 × TZEIORQ 43 was, however, the best hybrid under low-N conditions, and TZEIORQ 26 × TZEIORQ 47 was outstanding

for combined drought and low-N tolerance (Obeng-Bio, 2019).

Results of the combining ability study of PVA carotenoids and tryptophan revealed a preponderance of GCA effects over SCA for PVA and all measured carotenoids, indicating that superior hybrids could be produced by crossing parents with significant positive GCA effects for PVA. The hybrid TZEIORQ 29 × TZEIORQ 43, which recorded a PVA content of 9.78 μg g^{-1}, was among the hybrids identified with combined desirable agronomic performance under drought, low-N and optimal environments; it also had a moderate level of PVA contents. From this study, the outstanding hybrids, including TZEIRQ 29 × TZEIORQ 43, should be further tested to confirm consistency of performance and be commercialized in SSA to combat VAD and protein energy malnutrition in the sub-region.

The PVA allele-specific marker crtRB1-3′TE was identified as the relatively polymorphic marker and was highly consistent with the KASP SNP (snpZM0015). The two markers identified eight inbred lines harbouring the favourable alleles of the *crtRB1* functional gene. These inbreds could serve as donor parents of favourable alleles for the *crtRB1* gene. Despite the moderate to high PVA contents of TZEIORQ 29 and TZEIORQ 55, they were not validated as possessing the favourable alleles of *crtRB1* and *LcyE* genes, implying that other genes could be responsible for the increased levels of PVA in these inbreds. Moreover, the preponderance of additive genetic effects over the non-additive effects in the inheritance of PVA accumulation in the entire set of inbreds, and the recorded significant positive GCA male and female effects for PVA levels for TZEIORQ 29, indicated that TZEIORQ 55 and TZEIORQ 29 could contribute favourable alleles other than those of *crtRB1* and *LcyE* for the improvement of PVA concentrations in hybrids, synthetics and for the development of early PVA-QPM populations for extracting outstanding PVA inbred lines for hybrid development.

The preponderance of GCA (GCA-male + GCA-female) effects over SCA effects for grain yield and most agronomic traits under drought, low-N, optimal and across environments indicated that additive gene effects were more important than non-additive gene effects and that GCA largely contributed to the inheritance of the traits measured for the 96 early PVA-QPM hybrids evaluated (Obeng-Bio, 2019). This result therefore suggested that superior hybrids could be developed by crossing the parents with high significant and positive GCA effects (Baker, 1978; Badu-Apraku *et al*., 2013b). In contrast, another study conducted in IITA by Tchala (2019) revealed dominance genetic effects to be the primary type of effects conditioning most of the carotenoid traits in the extra-early set of parental inbreds.

In another study (Obeng-Bio, 2019), 54 early-maturing PVA-QPM single-cross hybrids generated from 18 inbred lines plus a hybrid check were phenotyped for carotenoids and tryptophan contents to determine the type of gene action conditioning the accumulation of PVA and tryptophan contents and to identify superior hybrids that combined elevated levels of PVA carotenoids and tryptophan with drought and low-N tolerance. PVA functional markers were also used to identify inbred lines harbouring the functional *crtRB1* and *lcyE* genes to serve as donor parents of the favourable alleles. Results revealed genetic variation for PVA carotenoids and tryptophan among the hybrids, which resulted in selection gains. It was concluded that carotenoid and tryptophan traits could be easily transferred from parental lines to progenies. The results also showed that, for the same hybrid, there could be significant differences in repeated samples for carotenoids and tryptophan, emphasizing the importance of precision and replication for accurate quantification. Additive gene action was more important than non-additive gene action in the inheritance of all carotenoids and tryptophan, and GCA was the major contributor to the heritable variation in carotenoids and tryptophan of the early PVA-QPM hybrids studied. It was found that cytoplasmic genes did not have a significant influence on the inheritance of carotenoids and tryptophan in the studied inbred lines. Inbred line TZEIORQ 29, with significant positive GCA (GCAm and GCAf) effects for PVA and its component carotenoids, could be exploited for the PVA favourable alleles in the development of high-PVA hybrids and synthetics, and for the improvement of the early PVA-QPM inbred population. Also, inbred TZEIORQ 13 displayed a significant positive GCA-female effect for PVA and β-carotene, indicating that it could be useful as a female parent for breeding for high-PVA maize. The moderate

range of PVA contents observed for the hybrids suggested that there was the need to introgress favourable PVA alleles from sources, such as the IITA extra-early inbred lines, TZEEIOR 202 and TZEEIOR 205, to improve the tropically adapted early PVA and PVA-QPM inbred populations and to facilitate the accumulation of PVA in available hybrids.

The five most outstanding early hybrids (TZEIORQ 29 × TZEIORQ 40, TZEIORQ 29 × TZEIORQ 43, TZEIORQ 29 × TZEIORQ 24, TZEIORQ 20 × TZEIORQ 29 and TZEIORQ 6 × TZEIORQ 29) had moderately high PVA levels and should be further tested and commercialized in SSA to combat protein energy malnutrition and VAD. There was a significant positive correlation between PVA carotenoids and grain yield of the PVA-QPM hybrids, indicating that simultaneous increases in accumulation of PVA and other carotenoids might be effectively accomplished without compromising grain yield potential of the hybrids.

In the PVA candidate genes validation study involving the PVA-QPM inbred lines, crtRB1-3′TE was the most polymorphic functional marker identified in these eight inbred lines, TZEIORQ 10, TZEIORQ 12, TZEIORQ 13, TZEIORQ 14, TZEIORQ 15, TZEIORQ 16, TZEIORQ 17 and TZEI 129, which indicated that they harboured the favourable alleles of the *crtRB1* functional gene. The KASP SNP, snpZM0015, was consistent with the results of the PCR-based markers. However, the crtRB1-5′TE and lcyE-5′TE did not amplify any of the inbreds. The eight inbreds identified could serve as donor parents of favourable alleles of the *crtRB1* gene. Information on the functional PVA genes and the phenotyping results of the carotenoids suggested that TZEIORQ 55 and TZEIORQ 29 could be sources of favourable alleles other than those of *crtRB1* and *LcyE* for the improvement of PVA concentrations in available hybrids, synthetics and the populations to be derived from the outstanding PVA-QPM inbred lines identified in the present study.

In the study, TZEIORQ 42 × TZEIORQ 20 ranked third among the 15 best hybrids and had its inbreds placed in the same heterotic group (group 4) (Obeng-Bio, 2019). The parental lines of the hybrid recorded significant ($P < 0.05$) positive GCA-male and female effects for GY. It had a relatively good grain-yielding ability under stress conditions to qualify as a seed parent in successful three-way and double-cross hybrids for high seed production. The single-cross hybrid, TZEIORQ 42 × TZEIORQ 20, was therefore, identified as a potential single-cross hybrid tester.

In breeding for improved levels of PVA in maize, many researchers have identified and used different molecular markers linked to PVA carotenoids (Harjes *et al.*, 2008; Yan *et al.*, 2010; Fu *et al.*, 2013; Sagare *et al.*, 2015b). Also, it has been reported that the presence of allele 1 (favourable allele) of *crtRB*1-3′TE could bring about a two- to ten-fold increase in kernel β-carotene concentration in maize (Babu *et al.*, 2013; Sagare *et al.*, 2015b). Similarly, in a marker–trait association study of functional gene markers for PVA levels across the tropical yellow maize inbred lines, Azmach *et al.* (2013) showed that the functional DNA markers *crtRB*1-3′TE and *crtRB*1-5′TE were polymorphic and strongly associated with PVA content across the tropical yellow maize lines tested. In another study in IITA, 76 PVA-QPM inbred lines plus four checks were screened by Oyekale (2019) for the presence of markers linked to either favourable or non-favourable β-carotene alleles. Allele-specific primers (*crtRB*1-3′TE and *crtRB*1-5′TE) were used to characterize the lines for the presence of markers linked to both (or either of the) alleles. The results were slightly different from those reported by earlier investigators in that only *crtRB*1-3′TE was polymorphic in the tropical extra-early PVA-QPM inbred lines screened with the functional DNA PVA markers, *crtRB*1-3′TE and *crtRB*1-5′TE. The differences in the results of this study and the earlier reports might be attributed to the differences in the genetic materials used in different studies. Furthermore, results of the HPLC indicated that 86% of the inbred lines harbouring the favourable allele of *crtRB*1-3′TE had high levels of β-carotene and total PVA levels greater than the mean PVA (6.2 μg g^{-1}) of all the extra-early PVA-QPM inbred lines analysed. This suggested that the marker is tightly linked to β-carotene with direct influence on the levels of total PVA in the inbreds. This result is consistent with the findings of many other investigators, who reported a strong association between the marker and the PVA carotenoid (Yan *et al.*, 2010; Babu *et al.*, 2013; Sagare *et al.*, 2015b). Although the inbred line (TZEEIORQ 54A) with the highest level of β-carotene eventually had the

highest level of total PVA (11 μg g^{-1}) of the inbred lines used for the study, a few of the inbreds, without the marker, had moderate levels of PVA relative to the HarvestPlus target of 15 μg g^{-1} (Bouis and Saltzman, 2017). This suggested that other PVA carotenoids, apart from β-carotene, might be more active in the inbred lines.

The HGCAMT method classified the extra-early PVA-QPM inbred lines into three heterotic groups and inbreds TZEEIORQ 61 and TZEEIORQ 5 were identified as testers for the inbreds across the research conditions. Also, inbreds TZEEIORQ 49 and TZEEIORQ 55, with combined *Striga* resistance and low-N tolerance as well as high levels of PVA, lysine and tryptophan, could be recombined to develop a population from which superior inbreds could be extracted. Equally, the extra-early PVA-QPM hybrids TZEEIORQ 61 × TZEEIORQ 49, TZEEIORQ 49 × TZEEIORQ 75 and TZEEIORQ 55 × TZEEIORQ 26 should be tested further for commercialization in SSA.

Stepwise regression analyses of all the PVA carotenoids on the total PVA levels in the extra-early PVA-QPM inbred lines revealed that β-cryptoxanthin (with half the vitamin A activity of β-carotene) contributed significantly (partial R^2 = 0.18) to the increased levels of total PVA in those inbred lines without the favourable allele of *crtRB*1-3′TE. This indicated that the allele-specific PVA marker *crtRB*1-3′TE was strongly linked to β-carotene with a resultant increase in PVA and that beta-cryptoxanthin was associated with the increased levels of PVA observed in some inbred lines that lacked the favourable PVA allele. Beta-cryptoxanthin equally might have made significant contribution to total PVA levels in the inbred lines possessing the favourable alleles. This result supported the findings of Venado *et al*. (2017), who showed a strong positive correlation (r = 0.70) between β-cryptoxanthin and PVA concentration in maize. The authors identified lycopene beta cyclase (*lcyB*) as the candidate gene associated with increased levels of β-cryptoxanthin in maize. Although Venado *et al*. (2017) used different genetic materials, the range of PVA levels (3.01–11.90 μg g^{-1}) and the average β-cryptoxanthin (4.23 μg g^{-1}) reported in their study were comparable to those obtained in the IITA study (PVA levels = 2.21–11.00 μg g^{-1}, average β-cryptoxanthin = 5.25 μg g^{-1}).

It is striking that in our studies at IITA, the levels of PVA were highest in the inbred lines with relatively deep orange kernels. This observation corroborated the results of many workers, who reported that high levels of total carotenoids, and slightly more PVA content, could be achieved when visual score for kernel colour was used in breeding for PVA-rich maize (Chandler *et al*., 2013; Venado *et al*., 2017). However, in their earlier report, Safawo *et al*. (2010) advocated other more efficient means of quantifying β-carotene in maize grains than kernel colour, following their observation that there was low correlation between visual grain colour and total carotenoid (R^2 = 0.184) as well as β-carotene content (R^2 = 0.033) of the 64 maize inbred lines evaluated in their study. Similarly, Muthusamy *et al*. (2015) reported that visual selection for kernel colour will be misleading in selecting PVA-rich genotypes, although it could improve the levels of non-PVA carotenoids, i.e. lutein and zeaxanthin.

The levels of tryptophan and lysine were at least 0.07 and 0.35%, respectively, in the following IITA inbred lines: TZEEIORQ 72, TZEEIORQ 74, TZEEIORQ 22 and TZEEIORQ 55, suggesting that the lines could be regarded as QPM inbred lines (Krivanek *et al*., 2007; Vivek *et al*., 2008; Tandzi *et al*., 2017). These results were consistent with the findings of Kostadinovic *et al*. (2016). Generally, inbred lines TZEEIORQ 58, TZEEIORQ 55, TZEEIORQ 5, TZEEIORQ 52, TZEEIORQ 57, TZEEIORQ 62, TZEEIORQ 72, TZEEIORQ 59 and TZEEIORQ 54 (with some of the highest levels of PVA, tryptophan and lysine) exhibited differential responses to both *Striga* infestation and low-N conditions. The first six top-yielding inbred lines were resistant/tolerant to both *Striga* and low-N stresses, whereas the last three were susceptible to the two stresses. The first six lines can be used to develop *Striga*-resistant/tolerant and low-N-tolerant extra-early-maturing PVA-QPM hybrids/varieties or populations for WCA (Oyekale, 2019).

The allele-specific DNA marker crtRB1-3′TE identified TZEEIORQ 54, TZEEIORQ 58, TZEEIORQ 55, TZEEIORQ 52, TZEEIORQ 57, TZEEIORQ 62, TZEEIORQ 51, TZEEIORQ 50, TZEEIORQ 60, TZEEIORQ 53, TZEEIORQ 49, TZEEIORQ 63, TZEEIORQ 64 and TZEEIORQ 73 as inbred lines

possessing favourable alleles for elevated levels of PVA. Also, biochemical analysis revealed that PVA levels ranged from 2.21 μg g^{-1} for TZEEIORQ 27 to 11.00 μg g^{-1} for TZEEIORQ 54, whereas tryptophan levels varied from 0.03% for TZEEIORQ 53 to 0.08% for TZEEIORQ 72 and lysine from 0.19% for TZEEIORQ 50 to 0.39% for TZEEIORQ 74.

Interrelationships among traits of extra-early provitamin A maize hybrids under drought and *Striga*-infested environments

Index selection has proved to be an effective selection method, as the trait of interest is usually selected along with other secondary traits influencing the observable expression of the target trait. Thus, information on inter-trait relationships guides the choice of traits a breeder would consider for inclusion in a selection index. Studies were conducted at IITA to investigate the correlations among grain yield and other agronomic traits of PVA maize hybrids and to determine the causal relationships among the PVA levels of selected hybrids, mid-parent PVA levels and other agronomic traits under managed drought and *Striga*-infested environments. One hundred and ninety diallel crosses developed from 20 PVA inbreds plus six checks were evaluated under drought and *Striga*-infested environments in Nigeria, 2015–2017. Grain yield and other agronomic traits of the hybrids were subjected to correlation analysis. Furthermore, the PVA content of 14 selected hybrids was determined along with those of the corresponding parental lines. Causal relationships among the hybrid PVA levels, mid-parent PVA levels and other traits were illustrated using stepwise regression and path analyses. Results revealed significant positive correlations between grain yield and other traits such as plant and ear heights, root lodging, ear rot and ears per plant under drought, whereas grain yield had significant negative correlations with days to anthesis and silking, stalk lodging, husk cover, plant and ear aspects and stay-green characteristic (Table 17.11).

Under *Striga*-infested environments, grain yield correlated positively with plant and ear heights and ears per plant but negatively with days to anthesis and silking, anthesis-silking interval, *Striga* damage at 8 and 10 weeks after planting, husk cover and ear aspect (Table 17.12). These results are to a large extent consistent with the findings of previous researchers under drought and *Striga*-infested environments. Sequential path analysis revealed mid-parent PVA as the single primary trait accounting for about 93% of the variation in the PVA levels of the hybrids under drought (Fig. 17.6). However, mid-parent PVA and root lodging were the primary traits influencing the PVA levels of the hybrids under *Striga*-infested environments. About 96% of the variation could be attributed to these traits (Fig. 17.7). Yield was identified as fifth- and third-order traits under drought and *Striga*-infested environments, respectively, suggesting that PVA levels of hybrids were independent of yield performance of the hybrids. Thus, simultaneous selection for high grain yield and elevated PVA levels would suffice.

Summary and Conclusions

Maize, a major staple food crop widely consumed in Africa, is deficient in nutritional quality, including two amino acids (lysine and tryptophan), minerals and vitamins, one of which is vitamin A. However, it contains low levels of the two amino acids, minerals (such as zinc and iron) and PVA. As a result of the existence of genetic variability for the quality composition of the kernels, two international agricultural research centres, specifically CIMMYT and IITA, supported by the HarvestPlus Challenge Programme, have been working to enhance the levels of nutritional quality traits of maize (hybrids and OPVs) to be released to farmers in SSA. Also, because maize production in SSA is greatly constrained by many abiotic (such as drought, heat and low-N) and biotic (including *S. hermonthica* infestation, diseases and insect pests) stress factors, development and deployment of high quality OPVs and hybrids with tolerance or resistance to multiple stresses that will simultaneously mitigate the problems of malnutrition and food insecurity in SSA, have been the main focus of the breeding programme. The maize germplasm available to the breeders contains a large number of inbred lines, OPVs and hybrids, into which tolerance of some or all of the stresses and/or quality protein traits have been incorporated.

Table 17.11. Correlations among grain yield and other agronomic traits[a] of diallel crosses of extra-early inbreds under managed drought stress at Ikenne during the 2015/16 and 2016/17 dry seasons.

	DA	DS	ASI	PHT	EHT	RL	SL	HUSK	PASP	EASP	EROT	STGR	EPP
YIELD	−0.60[c]	−0.70[c]	−0.12	0.35[c]	0.37[c]	0.16[c]	−0.29[c]	−0.67[c]	−0.82[c]	−0.87[c]	0.44[c]	−0.55[c]	0.78[c]
DA		0.79[c]	−0.38[c]	−0.25[c]	−0.28[c]	−0.02	0.25[c]	0.44[c]	0.58[c]	0.61[c]	−0.36[c]	0.38[c]	−0.64[c]
DS			0.27[c]	−0.26[c]	−0.32[c]	−0.04	0.22[c]	0.47[c]	0.60[c]	0.66[c]	−0.42[c]	0.38[c]	−0.68[c]
ASI				0.02	−0.04	−0.01	−0.05	0.02	−0.01	0.03	−0.06	−0.02	−0.02
PHT					0.70[c]	−0.04	−0.17[c]	−0.44[c]	−0.33[c]	−0.32[c]	0.18[c]	−0.14[b]	0.28[c]
EHT						0.02	−0.08	−0.28[c]	−0.29[c]	−0.29[c]	0.33[c]	−0.08	0.26[c]
RL							0.07	−0.04	−0.08	−0.10	0.07	0.00	0.12
SL								0.34[c]	0.36[c]	0.34[c]	−0.11	0.42[c]	−0.28[c]
HUSK									0.77[c]	0.77[c]	−0.36[c]	0.56[c]	−0.69[c]
PASP										0.85[c]	−0.47[c]	0.65[c]	−0.79[c]
EASP											−0.50[c]	0.61[c]	−0.84[c]
EROT												−0.21[c]	0.63[c]
STGR													−0.59[c]

[a]DA, days to 50% anthesis; DS, days to 50% silking; ASI, anthesis–silking interval; PHT, plant height; EHT, ear height; RL, root lodging; SL, stalk lodging; HUSK, husk cover; PASP, plant aspect; EASP, ear aspect; EROT, ear rot; STGR, stay-green characteristic; and EPP, ears per plant.

[b, c]Significant at 5% and 1% probability level, respectively.

Table 17.12. Correlations among grain yield and other agronomic traits[a] of diallel crosses of extra-early inbreds under artificial *Striga* infestation at Mokwa, during the 2016 and 2017 growing seasons.

	DA	DS	ASI	PHT	EHT	SDR1	SDR2	ESP1	ESP2	RL	SL	HC	EASP	EROT	EPP
YIELD	−0.40[c]	−0.55[c]	−0.44[c]	0.33[c]	0.21[c]	−0.76[c]	−0.78[c]	−0.11	−0.05	−0.11	−0.08	−0.77[c]	−0.88[c]	−0.01	0.78[c]
DA		0.83[c]	0.28[c]	−0.08	−0.09	0.36[c]	0.25[c]	−0.16[c]	−0.20[c]	0.00	−0.19[c]	0.31[c]	0.41[c]	−0.08	−0.41[c]
DS			0.68[c]	−0.15[c]	−0.13	0.52[c]	0.41[c]	−0.14[b]	−0.20[c]	0.03	−0.18[c]	0.42[c]	0.58[c]	−0.13	−0.56[c]
ASI				−0.20[c]	−0.17[c]	0.43[c]	0.38[c]	0.04	−0.06	0.02	−0.12	0.33[c]	0.49[c]	−0.13	−0.42[c]
PHT					0.69[c]	−0.45[c]	−0.32[c]	−0.04	−0.05	−0.01	0.03	−0.35[c]	−0.35[c]	0.04	0.29[c]
EHT						−0.23[c]	−0.13	0.00	0.06	0.01	0.20[c]	−0.13	−0.24[c]	0.04	0.23[c]
SDR1							0.75[c]	0.15[c]	0.06	0.07	0.14[b]	0.83[c]	0.69[c]	−0.01	−0.61[c]
SDR2								0.18[c]	0.20[c]	0.11	0.19[c]	0.85[c]	0.69[c]	0.04	−0.64[c]
ESP1									0.56[c]	0.02	0.23[c]	0.16[c]	0.02	0.18[c]	0.03
ESP2										−0.03	0.26[c]	0.17[c]	−0.08	0.08	0.06
RL											0.32[c]	0.12	0.06	0.08	0.00
SL												0.20[c]	0.00	0.08	0.11
HC													0.68[c]	0.07	−0.61[c]
EASP														−0.01	−0.77[c]
EROT															0.09

[a]DA, days to 50% anthesis; DS, days to 50% silking; ASI, anthesis–silking interval; PHT, plant height; EHT, ear height; SDR1 and SDR2, *Striga* damage (8 and 10 weeks after planting [WAP]), respectively; ESP1 and ESP2, emerged *Striga* plants (8 and 10 WAP); RL, root lodging; SL, stalk lodging; HC, husk cover; EASP, ear aspect; EROT, ear rot; and EPP, ears per plant.

[b,c]Significant at 5% and 1% probability level, respectively.

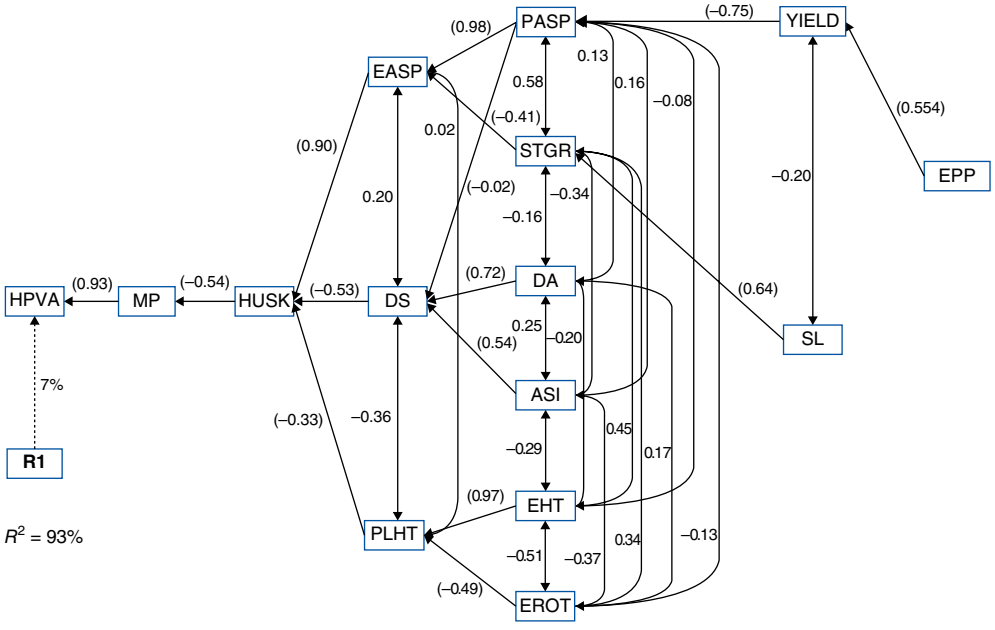


Fig. 17.6. Path analysis model diagram showing causal relationships of hybrid provitamin A (PVA) levels, mid-parent PVA levels and other measured traits of PVA diallel crosses evaluated under managed drought stress at Ikenne during the 2015/16 and 2016/17 dry seasons. Bold value (R1) is the residual effect; values in parenthesis are direct path coefficients and other values are correlation coefficients. ASI, anthesis–silking interval; EHT, ear height; EROT, ear rot; DA, days to 50% anthesis; DS, days to 50% silking; EASP, ear aspect; EPP, ears per plant; HPVA, hybrid provitamin A; HUSK, husk cover; MP, mid-parent provitamin A; PASP, plant aspect; PHT, plant height; STGR, stay-green characteristic; and SL, stalk lodging.

The materials, along with those obtained from other sources, such as CIMMYT, were screened by IITA maize scientists for PVA and were subjected to quantitative genetic studies and molecular approaches for PVA enhancement. Multiple stress-tolerant inbred lines have been used to develop early and extra-early OPVs and hybrids with high-quality protein and PVA for release to farmers of SSA. Although the HarvestPlus Challenge Programme has established 15 μg g^{-1} as the breeding target for PVA in maize and more than 40 initial PVA maize synthetics, single-cross and three-way hybrids have been released in the DRC, Ghana, Malawi, Mali, Nigeria, Rwanda, Tanzania, Zambia and Zimbabwe, only a few of the released PVA maize hybrids have attained this level in SSA; none of them has a quality protein background.

Literally thousands of entries with tolerance/resistance to multiple stresses and, in many cases, with a quality protein background (i.e. QPM) were screened for PVA content and subjected to intensive selection while maintaining farmers' desired agronomic traits, including high grain yield. Extensive quantitative genetic studies were conducted to quantify the genetic variability for PVA and determine its mode of inheritance, heritability, genotype × environment interaction, responses to recurrent selection and inter-relationships among agronomic traits, using sequential path analyses. Furthermore, the discovery of increased nutrition in yellow maize grain led to selection of pigmented grain as a desirable quality trait. Maize kernel colour and MAS have been employed in breeding PVA-rich maize. Genetic diversity has been estimated from various types of molecular markers, including RFLPs, AFLPs, SSRs, SNPs and DArT, which detect all types of DNA variations, including single-base changes and small insertions and deletions. The DArTseq markers have been found to be more efficient

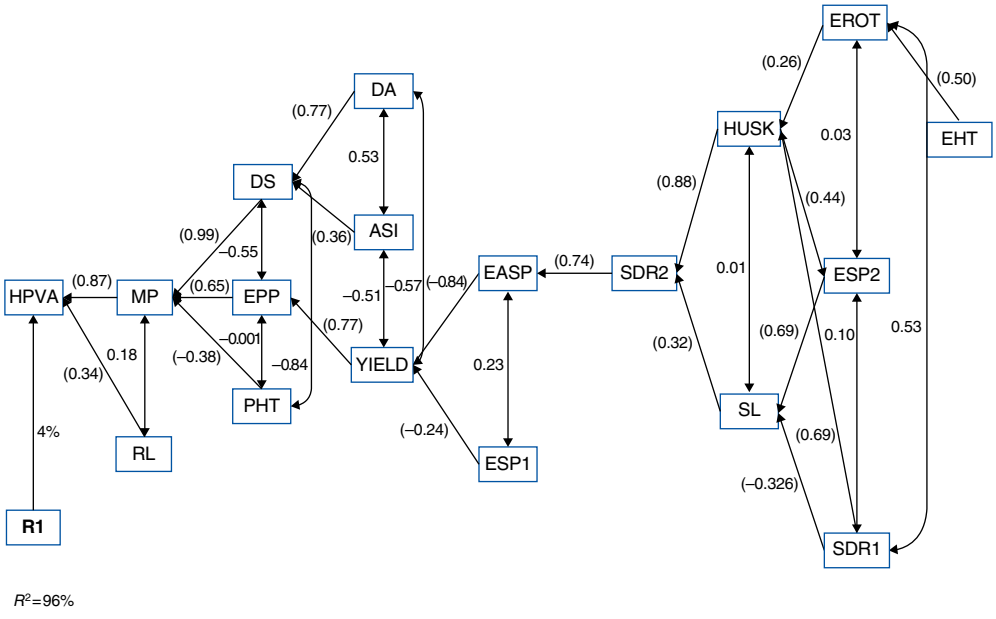


Fig. 17.7. Path analysis model diagram showing causal relationships of hybrid provitamin A (PVA) levels, mid-parent PVA levels and other measured traits of extra-early PVA diallel crosses evaluated under artificial *Striga* infestation at Mokwa, during the 2016 and 2017 growing seasons. Bold value (R1) is the residual effect; values in parenthesis are direct path coefficients and other values are correlation coefficients. HPVA, hybrid provitamin A; MP, mid-parent provitamin A; ASI, anthesis–silking interval; DA, days to 50% anthesis; DS, days to 50% silking; EASP, ear aspect; EPP, ears per plant; ESP1 and ESP2, number of emerged *Striga* plants (8 and 10 WAP); HUSK, husk cover; SL, stalk lodging; EROT, ear rot; PHT, plant height; RL, root lodging; SDR1 and SDR2, *Striga* damage (8 and 10 WAP); and EHT, ear height.

than the HGCAMT grouping method in identifying heterotic groups. Using this approach, TZEIORQ 29 was found to be the best male or female tester, whereas TZEIORQ 24 was the best male tester. TZEIORQ 59 × TZEIORQ 11 was identified as the best single-cross tester across research environments. A study of the genetic diversity and the population structure of 110 early-maturing PVA maize inbred lines from the IITA maize improvement programme evaluated under drought, *Striga*-infested and optimal conditions in 2015 and 2016 in Nigeria, revealed significant differences among the lines, indicating that the lines were genetically distinct and selection progress can be expected. The genetic distance between the early PVA lines ranged from 0.03 to 0.45. The genetic distance estimates showed that the early-maturing PVA inbred lines could be useful for hybrid production, population improvement and, eventually, development of new lines. The dendrogram obtained with DArT marker data placed the lines into five heterotic groups. Using the UPGMA clustering method, the early-maturing inbred lines were grouped into five clusters, which reflected parental relationships. The information obtained from this study would provide better understanding of the genetic relationships among the early PVA lines.

Results of studies conducted at IITA with tropically adapted maize germplasm showed that the orange-endosperm maize genotypes exhibited significant differences for all the traits measured, with more than 70% of the total variation observed being attributable to β-carotene, and the mean β-carotene content varied from 0.45 to 2.18 μg g^{-1}. The orange colour source, Syn-Y-STR-34-1-1-1-1-2-1-B-B-B-B-B/NC354/SYN-Y-STR-34-1-1-1 (OR1), used to convert 2004 TZEE-Y STR C4 to an orange population from which the extra early PVA inbred lines were derived, carried none of the functional alleles of

the *crtRB1* gene. Only TZEEIOR 196 and TZdEEI 7 contained the favourable allele at the 3′TE of *lycE* locus while all of the other inbred lines contained the unfavourable allele at 5′TE of *crtRB1* locus. Multiple stress-tolerant extra-early maize inbred lines with PVA levels higher than the target of 15 µg g^{-1} established by the HarvestPlus Programme have been identified at IITA; these included TZEEIOR 202 (23.98 µg g^{-1}) and TZEEIOR 205 (22.58 µg g^{-1}). Furthermore, an early-maturing PVA-QPM inbred line, TZEIORQ 55 (15.38 ug g^{-1}), has been identified. The extra-early and early PVA inbred lines are presently serving as invaluable high-PVA genetic resources for developing high-PVA hybrids and introgressing PVA alleles into tropical breeding populations. Furthermore, crosses involving the high-PVA maize inbred lines TZEEIOR 202 and TZEEIOR 205 have resulted in the development of PVA hybrids TZEEIOR 197 × TZEEIOR 205 (20.1 µg g^{-1}) and TZEEIOR 202 × TZEEIOR 205 (22.7 µg g^{-1}), containing about double the amount of PVA of the commercial PVA hybrid check, TZEE-Y Pop STR C5 × TZEEI 58 (11.4 µg g^{-1}), which are candidates for release in SSA.

In a series of studies, S_3 lines with deep orange colour were selected and recombined to form four extra-early PVA varieties (2009 TZEE-OR1 STR, 2009 TZEE-OR2 STR, 2009 TZEE-OR1 STR QPM and 2009 TZEE-OR2 STR QPM), four early PVA varieties (2009 TZE-OR1 STR, 2009 TZE-OR2 STR, 2009 TZE-OR1 STR QPM and 2009 TZE-OR2 STR QPM), two normal-endosperm PVA varieties (2009 TZEE-OR1 STR and 2009 TZE-OR2 STR) and two PVA-QPM varieties (2009 TZEE-OR1 STR QPM and 2009 TZE-OR2 DT STR QPM). Many multi-stress-tolerant/resistant PVA inbred lines have been extracted from the populations. The inbred lines are presently being used in various genetic studies to (i) determine their combining abilities and heterotic patterns; (ii) classify them into heterotic groups; (iii) identify inbred and single-cross testers; and (iv) determine the performance and stability of the inbreds in hybrid combinations.

Results of many quantitative genetic studies of PVA, conducted in multi-environment trials (METs) by IITA and NARS scientists in WA indicated the presence of both additive and non-additive genetic variances with a preponderance of the additive component; several distinct heterotic groups with large, highly significant heterosis in cross combinations; high heritability estimates; and a desirable response to recurrent selection for improved PVA. Many inbred lines (such as TZEEIORQ 49, TZEEIORQ 55, TZEEIORQ 53 and TZEEIORQ 5) with combined *Striga* resistance and low-N tolerance as well as high levels of lysine, tryptophan and PVA levels much higher than the 15 µg g^{-1} initially recommended by the HarvestPlus Programme, have been identified as testers for single-cross hybrids, and several single-cross hybrids (such as TZEIORQ 59 × TZEIORQ 11 and TZEIORQ 11 × TZEIORQ 29) have been identified as testers for producing three-way hybrids. In addition, several hybrids (including TZEIORQ 29 × TZEIORQ 40, TZEIORQ 29 × TZEIORQ 43, TZEIORQ 29 × TZEIORQ 24, TZEIORQ 20 × TZEIORQ 29 and TZEIORQ 6 × TZEIORQ 29) are in the pipeline for release to SSA farmers. In one study, however, a negative trend in grain yield was observed in association with single-trait selection for enhanced PVA concentration for two of the three populations included in the study, contradicting the results of many previous studies, which had reported lack of correlation between grain yield and PVA concentrations in maize. It was hypothesized that the specific circumstances of the project may be responsible for this undesirable association. To overcome the constraints that may be caused by the contradictory results, if real, breeding for enhanced PVA concentrations in maize should simultaneously consider grain yield. It may be concluded that maize in SSA can be effectively subjected to genetic enhancement of PVA, along with other mineral components of the kernel and the plant traits for sustainable, high-quality food sufficiency to drastically reduce hunger and malnutrition.

Acknowledgements

The authors are grateful to Dr L. Konate and Ms O. Olatise for parts of their thesis findings reported in this chapter. We also appreciate the financial support of the STMA Project for the research activities presented in this chapter. The activities reported in this chapter were also supported by the IITA management, especially Dr Robert Asiedu, Director of the West Africa Hub, and Dr May-Guri Seathre, IITA Deputy Director General of R4D, which has contributed

significantly to the success of the early and extra-early maize programme at IITA. We acknowledge the support of the IITA Technical Staff of the maize improvement programme and national maize scientists of West and Central Africa for their collaboration. Finally, we are very grateful to Dr J. Toyinbo and Mr S. Adewale for their critical review of this chapter.

References

Acquaah, G. (2012) *Principles of Plant Genetics and Breeding*, 2nd edn. Wiley, Oxford, UK.

Adofo-Boateng, P. (2015) Development of high-yielding and stable maize (*Zea mays* L.) hybrids tolerant to low soil nitrogen. PhD dissertation, University of Ghana, Accra, Ghana. Available at: http://ugspace.ug.edu.gh (accessed 20 June 2018).

Aguayo, V.M. and Baker, S.K. (2005) Vitamin A deficiency and child survival in sub-Saharan Africa: A reappraisal of challenges and opportunities. *Food and Nutrition Bulletin* 26, 348–355.

Akinwale, R.O., Badu-Apraku, B., Fakorede, M.A.B. and Vroh, I. (2014) Heterotic grouping of tropical early-maturing maize inbred lines based on combining ability in *Striga*-infested and *Striga*-free environments and the use of SSR markers for genotyping. *Field Crops Research* 156, 48–62.

Al-Babili, S. and Beyer, P. (2005) Golden rice – five years on the road five years to go? *Trends in Plant Science* 10, 565–573.

Amegbor, I.K., Badu-Apraku, B. and Annor, B. (2017) Combining ability and heterotic patterns of extra-early maturing white maize inbreds with genes from *Zea diploperennis* under multiple environments. *Euphytica* 213(1), 24. DOI: 10.1007/s1068-1-016-1823-y.

Annor, B. and Badu-Apraku, B. (2016) Gene action controlling grain yield and other agronomic traits of extra-early quality protein maize under stress and non-stress conditions. *Euphytica* 212(2), 213–228.

Arango, J., Jourdan, M., Geoffriau, E., Beyer, P. and Welsch, R. (2014) Carotene hydroxylase activity determines the levels of both α-carotene and total carotenoids in orange carrots. *Plant Cell* 26, 2223–2233. DOI: 10.1105/tpc.113.122127.

Auldridge, M.E., McCarty, D.R. and Klee, H.J. (2006) Plant carotenoid cleavage oxygenases and their apocarotenoid products. *Current Opinion in Plant Biology* 9, 315–321.

Azmach, G., Gedil, M., Menkir, A. and Spillane, C. (2013) Marker-trait association analysis of functional gene markers for provitamin A levels across diverse tropical yellow maize inbred lines. *BMC Plant Biology* 13(1), 227. DOI: 10.1186/1471-2229-13-227.

Babu, R., Rojas, N. P., Gao, S., Yan, J. and Pixley, K. (2013) Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. *Theoretical and Applied Genetics* 126(2), 389–399.

Badu-Apraku, B., Akinwale, R.O., Franco J. and Oyekunle, M. (2012) Assessment of reliability of secondary traits in selecting for improved grain yield in drought and low-nitrogen environments. *Crop Science* 52, 2050–2062.

Badu-Apraku, B., Oyekunle, M., Akinwale, R.O. and Aderounmu. M. (2013a) Combining ability and genetic diversity of extra-early white maize inbreds under stress and nonstress environments. *Crop Science* 53, 9–26.

Badu-Apraku, B., Oyekunle, M., Fakorede, M.A.B., Vroh, I., Akinwale, R.O., *et al*. (2013b) Combining ability, heterotic patterns and genetic diversity of extra-early yellow inbreds under contrasting environments. *Euphytica* 192, 413–433.

Badu-Apraku, B., Annor, B., Oyekunle, M., Akinwale, R.O., Fakorede, M.A.B., *et al*. (2015a) Grouping of early maturing quality protein maize inbreds based on SNP markers and combining ability under multiple environments. *Field Crops Research* 183, 169–183.

Badu-Apraku, B., Annor, B., Oyekunle, M., Fakorede, M.A.B., Akinwale, R.O., *et al*. (2015b) Grouping of early maturing quality protein maize inbreds based on SNP markers and combining ability under multiple environments. *Field Crops Research* 183, 169–183.

Badu-Apraku, B., Fakorede. M.A.B., Talabi, A.O., Oyekunle, M., Akaogu, I.C., *et al*. (2016a) Gene action and heterotic groups of early white quality protein maize inbreds under multiple stress environments. *Crop Science* 56(1), 183–199.

Badu-Apraku, B., Fakorede, M.A.B., Gedil, M., Annor, B., Talabi, A.O., *et al*. (2016b) Heterotic patterns of IITA and CIMMYT early-maturing yellow maize inbreds under contrasting environments *Agronomy Journal* 108, 1–16.

Baker, R.J. (1978) Issues in diallel analysis. *Crop Science* 18, 533–536.
Bänziger, M., Edmeades, G.O. and Lafitte, H.R. (1999) Selection for drought tolerance increases maize yields over a range of N levels. *Crop Science* 39, 1035–1040.
Barata, C. and Carena, M. (2006) Classification of North Dakota maize inbred lines into heterotic groups based on molecular and testcross data. *Euphytica* 151, 339–349.
Betrán, J., Beck, D., Bänziger, M., and Edmeades, G.O. (2003) Genetic analysis of inbred and hybrid grain yield under stress and non-stress environments in tropical maize. *Crop Science* 43, 807–817.
Blessin, C.W., Brecher, J.D. and Dimler, R.J. (1963) Carotenoids of corn and sorghum V. Distribution of xanthophylls and carotenes in hand-dissected and dry-milled fractions of yellow dent corn. *Cereal Chemistry* 40, 582–586.
Bolaños, J. and Edmeades, G.O. (1993) Eight cycles of drought tolerance in lowland tropical maize: Response in grain yield, biomass, and radiation utilization. *Field Crops Research* 31, 233–252.
Bouis, H.E. and Saltzman, A. (2017) Improving nutrition through biofortification: A review of evidence from HarvestPlus, 2003 through 2016. *Global Food Security* 12, 49–58.
Bouis, H.E., Harris, S. and Lineback, D. (2002). Biotechnology-derived nutritious foods for developing countries: Needs, opportunities, and barriers. *Food Nutrition Bulletin* 23, 342–383.
Brenna, O.V. and Berardo, N. (2004) Application of near-infrared reflectance spectroscopy (NIRS) to the evaluation of carotenoids content in maize. *Journal of Agricultural and Food Chemistry* 52, 5577–5582.
Brunson, A.M. and Quackenbush, F.W. (1962) Breeding corn with high provitamin A in the grain. *Crop Science* 2, 344–347.
Buckner, B., Kelson, T.L. and Robertson, D.S. (1990) Cloning of the *y1* locus of maize, a gene involved in the biosynthesis of carotenoids. *Plant Cell* 2, 867–876.
Burt, A.J., Grainger, C.M., Smid, M.P., Shelp, B.J. and Lee, E.A. (2011) Allele mining of exotic maize germplasm to enhance macular carotenoids. *Crop Science* 51, 991–1004.
Chander, S., Guo, Y.Q., Yang, X.H., Zhang, J., Lu, X.Q., *et al*. (2008) Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. *Theoretical and Applied Genetics* 116, 223–233.
Chandler, K., Lipka, A.E., Owens, B.F., Li, H., Buckler, E.S., Rocheford, T., *et al*. (2013) Genetic analysis of visually scored orange kernel color in maize. *Crop Science* 53, 189.
Coors, J.G. (1999) Selection methodologies and heterosis. In: Coors, J.G. and Pandey S. (eds) *Genetics and Exploitation of Heterosis in Crops*. American Society of Agronomy, Inc., Madison, Wisconsin. pp. 225–245.
Cruz, V.M.V., Kilian A. and Dierig, D.A. (2013) Development of DArT marker platforms and genetic diversity assessment of US. Collection of the new oilseed crop Lesquerella and related species. *PLoS ONE* 8(5), e64062. DOI: 10.1371/journal.pone.e64062.
Dhliwayo, T., Palacios-Rojas, N., Crossa, J. and Pixley, K. V. (2014) Effects of S 1 recurrent selection for provitamin A carotenoid content for three open-pollinated maize cultivars. *Crop Science* 54(6), 2449–2460.
Egesel, C.O., Wong, J.C., Lambert, R.J. and Rocheford, T.R. (2003a) Combining ability of maize inbreds for carotenoids and tocopherols. *Crop Science* 43, 818–823.
Egesel, C.O., Wong, J.C., Lambert, R.J. and Rocheford, T.R. (2003b) Gene dosage effects on carotenoid concentration in maize grain. *Maydica* 48, 183–190.
Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., *et al*. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6, e19379. DOI: 10.1371/journal.pone.0019379.
Fan, X.M., Tan, J., Yang, J.Y. and Chen, H.M. (2004) Combining ability and heterotic grouping of ten temperate, subtropical and tropical quality protein maize inbreds. *Maydica* 49, 267–272.
Fan, X.M., Chen, H.M., Tan, J., Xu, C.X., Zhang, Y.M., *et al*. (2008) A new maize heterotic pattern between temperate and tropical germplasms. *Agronomy Journal* 100(4), 917–923.
Fan, X.M., Zhang, Y.M., Yao, W.H., Chen, H.M., Tan, J., *et al*. (2009) Classifying maize inbred lines into heterotic groups using a factorial mating design. *Agronomy Journal* 101, 106–112.
Food and Agriculture Organization of the United Nations (FAO) (1994) Medium-term prospects for agricultural commodities, projections to the year 2000. FAO Economic and Social Development Paper No. 120. FAO, Rome, Italy.
Ford, F.H. (2000) Inheritance of kernel color in corn: Explanations and investigations. *American Biology Teacher* 62, 181–188.
Forgey, W.M. (1974) Inheritance of isomers of vitamin E in *Zea mays* L. MSc Thesis, University of Illinois at Urbana-Champaign, Illinois.

Fu, Z., Chai, Y., Zhou, Y., Yang, X., Warburton, M. L., *et al*. (2013) Natural variation in the sequence of *PSY*1 and frequency of favorable polymorphisms among tropical and temperate maize germplasm. *Theoretical and Applied Genetics* 126, 923–935.

Galicia, L., Nurit, E., Rosales, A., and Palacios-Rojas, N. (2008) Total starch determination in maize grains using a modified assay from megazyme. Laboratory Protocols. Maize Nutrition Quality and Plant Tissue Analysis Laboratory. CIMMYT, Texcoco, Mexico, pp. 34–37.

Galicia, L., A. Miranda, M.G. Gutiérrez, O. Custodio, A. Rosales, *et al*. (2012) *Laboratorio de calidad nutricional de maíz y análisis de tejido vegetal: Protocolos de laboratorio 2012*. CIMMYT, Mexico, D.F. Available at: http://repository.cimmyt.org/xmlui/bitstream/handle/10883/1349/97125.pdf (accessed 26 December 2018).

Gama, J., Sylos, C. and Dufosse, L. (2005) Major carotenoid composition of Brazilian Valencia orange juice: Identification and quantification by HPLC. *Food Research International* 38, 899–903.

Gebremeskel, S., Garcia-Oliveira, A.L., Menkir, A., Adetimirin, V. (2018) Effectiveness of predictive markers for marker assisted selection of pro-vitamin A carotenoids in medium-late maturing maize (*Zea mays* L.) inbred lines. *Journal of Cereal Science* 79, 27–34.

Gibson, G. (2010) Hints of hidden heritability in GWAS. *Nature Genetics* 42(7), 558–560.

Griffing, B. (1956) Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Sciences* 9, 463–493.

Grogan, C.O., Blessin, C.W., Dimler, R.J. and Campbell, C.M. (1963) Parental influence on xanthophylls and carotenoids in corn. *Crop Science*, 3, 213–214.

Halilu, A.D. (2016) Genetics of carotenoids in tropical-adapted pro-vitamin A maize (*Zea mays* L.). PhD Dissertation, Ahmadu Bello University, Zaria, Nigeria. 140 pp.

Halilu, A.D., Ado, S.G., Aba, D.A. and Usman, I.S. (2016) Genetics of carotenoids for provitamin A biofortification in tropical-adapted maize. *The Crop Journal* 4(4), 313–322.

Hallauer, A.R. and Miranda. J.B. (1988) *Quantitative Genetics in Maize Breeding*, 2nd edition. Iowa State University Press, Ames, Iowa.

Hallauer, A.R. and Carena, M. J. (2009) Maize. In: Carena, M.J. (ed.) *Cereals*. Springer, New York, pp. 3–98.

Harjes, C.E., Rocheford, T., Ling, B., Brutnell, T.P., Kandianis, C.B., *et al*. (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* 319(5861), 330–333.

HarvestPlus (2004) *Micronutrient Malnutrition. Vitamin A. Fifth Report on the World Nutrition Situation*. UN SCN, March 2004. Available at: https://www.unscn.org/layout/modules/resources/files/rwns5.pdf (accessed 24 August 2018).

Hayman, B.I. (1954) The theory and analysis of diallel crosses. *Genetics* 39, 789–809.

Holland, J.B. (2001) Epistasis and plant breeding. *Plant Breeding Reviews* 21, 27–92. Available at: http://www4.ncsu.edu/~jholland/Pubs/Epistasis.pdf (accessed 25 December 2015).

Hosana, G. C., Alamerew, S., Tadesse, B. and Menamo, T. (2015) Test cross performance and combining ability of maize (*Zea mays* L.) inbred lines at Bako, Western Ethiopia. *Global Journal of Science Frontier Research* 15(4), 1–24.

Howe, J.A. and Tanumihardjo, S.A. (2006) Carotenoid-biofortified maize maintains adequate vitamin A status in Mongolian gerbils 1. *Journal of Nutrition* 136, 2562–2567.

Howitt, C.A. and Pogson, B.J. (2006) Carotenoid accumulation and function in seeds and non-green tissues. *Plant, Cell and Environment* 29(3), 435–445.

Hung, H.Y. and Holland, J.B. (2012) Diallel analysis of resistance to fusarium ear rot and fumonium contamination in maize. *Crop Science* 52, 2173–2181.

Islam, S.N. (2004) Survey of carotenoid variation and quantitative trait loci mapping for carotenoid and tocopherol variation in maize. Master's thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois.

Khush, G.S. (2002) The promise of biotechnology in addressing current nutritional problems in developing countries. *Food and Nutrition Bulletin* 23(4), 354–357.

Kim, S. K. and Adetimirin, V.O. (1997) Responses of tolerant and susceptible maize varieties to timing and rate of nitrogen under *Striga hermonthica* infestation. *Agronomy Journal* 89(1), 38–44.

Kiula, B.A., Lyimo, N.G. and Botha, A.M. (2008) Association between AFLP-based genetic distance and hybrid performance in tropical maize. *Plant Breeding* 127, 140–144.

Konate, L., Badu-Apraku, B. and Traoré, D. (2017) Combining ability and heterotic grouping of early maturing pro-vitamin A maize inbreds across *Striga* infested and optimal environments. *Journal of Agriculture and Environment for International Development* 111(1), 157–173.

Konstantinov, K. and Mladenović-Drinić, S. (2007) Molecular genetics: Step by step implementation in maize breeding. *Genetika* 39(2), 139–154.

Kostadinovic, M., Ignjatovic-Micic, D., Vancetovic, J., Ristic, D., Bozinovic, S., *et al*. (2016) Development of high tryptophan maize near isogenic lines adapted to temperate regions through marker assisted selection – impediments and benefits. *PLoS ONE* 11(12), e0167635.

Krivanek, A.F., Groote, H.D., Gunaratna, N.S., Diallo, A.O. and Friese, D. (2007) Breeding and disseminating quality protein maize (QPM) for Africa. *African Journal of Biotechnology* 6, 312–324.

Kurilich, A.C. and Juvik, J.A. (1999) Quantification of carotenoid and tocopherol antioxidants in *Zea mays*. *Journal of Agricultural and Food Chemistry* 47(5), 1948–1955.

Le, D.T., Chua, H.D. and Le, N.Q. (2016) Improving nutritional quality of plant proteins through genetic engineering. *Current Genomics* 17, 220–229.

Lee, M., Godshalk, E.B. Lamkey, K.R. and Woodman, W.W. (1989) Association of restriction fragment length polymorphism among maize inbreds with agronomic performance of their crosses. *Crop Science* 29, 1067–1071.

Liu, K., Goodman, M., Muse, S., Smith, J.S., Buckler E., *et al*. (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165(4), 2117–2128.

Lozano-Alejo, N., Carrillo, G.V., Pixley, K. and Palacios-Rojas, N. (2007) Physical properties and carotenoid content of maize kernels and its nixtamalized snacks. *Innovative Food Science and Emerging Technologies* 8(3), 385–389.

Melchinger, A.E. (1999) Genetic diversity and heterosis. In: J.G. Coors and S. Pandey (eds) *The Genetics and Exploitation of Heterosis in Crops*. American Society of Agronomy, Madison, Wisconsin, pp. 99–118.

Melchinger, A.E., Lee, M., Lamkey, K.R., Hallauer, A.R. and Woodman, W.L. (1990) Genetic diversity from restriction fragment length polymorphisms and heterosis for two diallel sets of maize inbreds. *Theoretical and Applied Genetics* 80, 488–496.

Menkir, A. and Maziya-Dixon, B. (2004) Influence of genotype and environment on β-carotene content of tropical yellow-endosperm maize genotypes. *Maydica* 49, 313–318.

Menkir, A., Melake-Berhan, A., Ingelbrecht, C.I. and Adepoju, A. (2004) Grouping of tropical mid-altitude maize inbred lines on the basis of yield data and molecular markers. *Theoretical and Applied Genetics* 108, 1582–1590.

Menkir, A., Liu W., White, W.S., Maziya-Dixon, B. and Rocheford, T. (2008) Carotenoid diversity in tropical-adapted yellow maize inbred lines. *Food Chemistry* 109, 521–529.

Menkir, A., Gedil M., Tanumihardjo, S., Adepoju, A. and Bossey, B. (2014) Carotenoid accumulation and agronomic performance of maize hybrids involving parental combinations from different marker-based groups. *Food Chemistry* 148, 131–137.

Menkir, A., Maziya-Dixon, B., Mengesha, W., Rocheford, T. and Alamu, E.O. (2017) Accruing genetic gain in pro-vitamin A enrichment from harnessing diverse maize germplasm. *Euphytica* 213(5), 105.

Mishra, P. and Singh, N.K. (2010) Spectrophotometric and TLC based characterization of kernel carotenoids in short duration maize. *Maydica* 55(2), 95–100.

Moll, R.H., Lonnquist, J.H., Fortuna, J.V. and Johnson, E.C. (1965) The relation of heterosis and genetic divergence in maize. *Genetics* 52, 139–144.

Muthayya, S., Hyu Rah, J., Sugimoto, J., Roos, F., Kraemer, K., *et al*. (2013) The global hidden hunger indices and maps: An advocacy tool for action. *PLoS ONE* 8(6), e67860. DOI: 10.1371/journal.journal.pone.0067860.

Muthusamy, V., Hossain, F., Thirunavukkarasu, N., Saha, S., Agrawal, P.K., *et al*. (2015) Genetic variability and inter-relationship of kernel carotenoids among indigenous and exotic maize (*Zea mays* L.) inbreds. *Cereal Research Communications* 43(4), 567–578.

Muzhingi, T., Yeum, K.J., Russell, R.M., Johnson, E.J., Qin J., *et al*. (2008) Determination of carotenoids in yellow maize, the effects of saponification and food preparations. *International Journal for Vitamin and Nutrition Research* 78(3), 112–120.

Nuss, E.T. and Tanumihardjo, S.A. (2011) Quality protein maize for Africa: Closing the protein inadequacy gap in vulnerable populations. *Advances in Nutrition* 2, 217–224.

Nyquist, W.E. and Baker, R.J. (1991) Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences* 10(3), 235–322.

Obeng-Bio. E. (2019) Genetic analysis of grain yield and other traits of early maturing pro-vitamin A-quality protein maize inbred lines in contrasting environments. PhD Thesis, West Africa Centre for Crop Improvement, University of Ghana, Legon, Accra, Ghana.

Olatise, O. (2018) Combining ability and heterotic grouping of extra-early provitamin A maize inbred lines under stress and non-stress conditions. PhD Thesis, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria.

Ortiz, D., Rocheford, T. and Ferruzzi, M.G. (2016) Influence of temperature and humidity on the stability of carotenoids in biofortified maize (*Zea mays* L.) genotypes during controlled postharvest storage. *Journal of Agricultural and Food Chemistry* 64(13), 2727–2736.

Ortiz-Monasterio, J.I., Palacios-Rojas, N., Meng, E., Pixley, K., Trethowan, R., (2007) Enhancing the mineral and vitamin content of wheat and maize through plant breeding. *Journal of Cereal Science* 46(3), 293–307.

Oyekale, A.S. (2019) Genetic analysis of *Striga* resistance and tolerance to low soil nitrogen in crosses of extra-early of provitamin A-quality protein maize (*Zea Mays* L.) inbred lines in Nigeria. PhD Thesis, Pan African University Institute of Life and Earth Sciences, Ibadan, Oyo State, Nigeria.

Pfeiffer, W.H., and McClafferty, B. (2007) HarvestPlus: Breeding crops for better nutrition. *Crop Science* 47, S88–S105.

Pixley, K., Rojas, N.P., Babu, R., Mutale, R., Surles, R. and Simpungwe, E. (2013) Biofortification of maize with provitamin A carotenoids. In: Tanumihardjo, S.A. (ed.) *Carotenoids and Human Health*. Springer, New York, pp. 271–292.

Quackenbush, F.W. (1963) Corn carotenoids: Effects of temperature and moisture on losses during storage. *Cereal Chemistry* 40, 266–269.

Quackenbush, F.W., Firch, J.G., Rabourn, W.J., McQuistan, M., Petzold, W.N., *et al*. (1961) Composition of corn, analysis of carotenoids in corn grain. *Journal of Agricultural and Food Chemistry* 9(2), 132–135.

Quackenbush, F.W., Firch, J.G., Brunson, A.M. and House, L.R. (1966) Carotenoid, oil and tocopherol content of corn inbreds. *Cereal Chemistry* 40, 251–259.

Safawo, T., Senthil, N., Raveendran, M., Vellaikumar, S., Ganesan, K. N., Nallathambi, G., *et al*. (2010) Exploitation of natural variability in maize for β-carotene content using HPLC and gene specific markers. *Electronic Journal of Plant Breeding* 1(4), 548–555.

Sagare, D.B., Reddy, S.S., Shetti, P. and Surender, M. (2015a) Enhancing provitamin A of maize using functional gene markers. *International Journal of Advanced Biotechnology and Research* 6(1), 86–95.

Sagare, D.B., Shetti, P., Reddy, S.S., Surender, M. and Pradeep, T. (2015b) Identification of β-carotene rich maize inbreds using PCR-based assay for *crtRB*1-3′TE allele. *International Journal of Science and Nature* 6(3), 441–443.

Saltzman, A., Birol, E., Bouis, H. E., Boy, E., De Moura, F. F., *et al*. (2013) Biofortification: Progress toward a more nourishing future. *Global Food Security* 2(1), 9–17.

Sanchez-Sevilla, J.F., Horvath, A., Botella, M.A., Gaston, A., Folta, K., *et al*. (2015) Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria x ananassa*). *PLoS ONE* 10(12), e0144960.

Sansaloni, C.P., Petroli, C.D., Carling, J., Hudson, C.J., Steane, D.A., *et al*. (2010) A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. *Plant Methods* 6(1), 16.

Semagn, K., Magorokosho, C., Vivek, B.S., Makumbi, D., Beyene, Y., *et al*. (2012) Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. *BMC Genomics* 13, 113.

Semagn, K., Babu, R., Hearne, S. and Olsen, M. (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Molecular Breeding* 33(1), 1–14.

Smith, J.S.C. and Smith, O.S. (1992) Fingerprinting crop varieties. *Advances in Agronomy* 47, 85–140.

Smith, O.S., Smith, J.S.C., Bowen, S.L., Tenborg, R.A., and Wall, S.J. (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield, heterosis and RFLPs. *Theoretical and Applied Genetics* 80, 833–840.

Sofi, P.A., Wani, A.S., Rather, A.G. and Wani, H.S. (2009) Quality protein maize (QPM): Genetic manipulation for the nutritional fortification of maize. Review article. *Journal of Plant Breeding and Crop Science* 1, 244–253.

Sprague, G.F. and Tatum, L.A. (1942) General vs. Specific Combining Ability in Single Crosses of Corn 1. *Agronomy Journal* 34(10), 923–932.

Sserumaga, J.P., Makumbi, D., Ji, H., Njoroge, K., Muthomi, J.W., *et al*. (2014) Molecular characterization of tropical maize inbred lines using microsatellite DNA markers. *Maydica* 59, 267–274.

Stranger, B.E., Stahl, E.A. and Raj, T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187(2), 367–383.

Sughroue, J.R. (1995) Proper analysis of the diallel mating design. PhD thesis, Iowa State University, Ames, Iowa. 95 pp.

Suwarno, W.B., Pixley, K.V., Palacios-Rojas, N., Kaeppler, S.M. and Babu, R. (2014) Formation of heterotic groups and understanding genetic effects in a provitamin A biofortified maize breeding program. *Crop Science* 54(1), 14–24.

Tandzi, L.N., Mutengwa, C.S., Eddy L.M., Ngonkeu, E.D.L., Woïn, N., *et al*. (2017) Breeding for quality protein maize (QPM) varieties: A review. *Agronomy* 7, 80. DOI: 10.3390/agronomy7040080.

Tchala, N. (2019) Molecular characterization and heterosis of extra-early maturing orange maize inbred lines under contrasting environments. PhD Thesis, West African Center for Crop Improvement, University of Ghana, Legon, Accra, Ghana.

US Institute of Medicine (2001) Vitamin A. In: *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. National Academy Press, Washington DC, pp. 82–161. Available at: https://www.nap.edu/catalog/10026/dietary-reference-intakes-for-vitamin-a-vitamin-k-arsenic-boron-chromium-copper-iodine-iron-manganese-molybdenum-nickel-silicon-vanadium-and-zinc (accessed 6 January 2013).

Vallabhaneni, R., Gallagher, C.E., Licciardello, N., Cuttriss, A.J., Quinlan, R.F., *et al*. (2009) Metabolite sorting of a germplasm collection reveals the Hydroxylase3 locus as a new target for maize provitamin A biofortification. *Plant Physiology* 151, 1635–1645.

Vallabhaneni, R., Bradbury, L.M. and Wurtzel, E.T. (2010) The carotenoid dioxygenase gene family in maize, sorghum and rice. *Archives of Biochemistry and Biophysics* 504(1), 104–111.

Venado, R.E., Owens, B.F., Ortiz, D. Lawson, T., Mateos-Hernandez, M., *et al*. (2017) Genetic analysis of provitamin A carotenoid β-cryptoxanthin concentration and relationship with other carotenoids in maize grain (*Zea mays* L.). *Molecular Breeding* 37(10), 127.

Vivek, B.S., Krivanek, A.F., Palacios-Rojas, N., Twumasi-Afriyie, S. and. Diallo, A.O (2008) *Breeding Quality Protein Maize (QPM): Protocols for Developing QPM Cultivars*. CIMMYT, Texcoco, Mexico.

Weber, E.J. (1987) Carotenoids and tocols of corn grain determined by HPLC. *Journal of the American Oil Chemists Society* 64, 1129–1134.

West, K.P. Jr. and Darnton-Hill, I. (2008) Vitamin A deficiency. In: Semba R.D. and Bloem M.W. (eds) *Nutrition and Health in Developing Countries*. Humana Press, Totowa, New Jersey, pp. 377–434.

Wong, J.C., Egesel, C.O., Kurilich, A.C., Rocheford, T.R., Lambert, R.J., *et al*. (1998) Genetic variation on maize for vitamin A and vitamin E. In: *Proceedings of the 34th Annual Illinois Corn Breeders' School*. University of Illinois, Urbana-Champaign, Illinois, pp. 191–202.

Wong, J.C., Lambert, R.J., Wurtzel, E.T. and Rocheford, T.R. (2004) QTL and candidate genes phytoene synthase and z-carotene desaturase associated with the accumulation of carotenoids in maize. *Theoretical and Applied Genetics* 108, 349–359.

World Health Organization (2009) *Global Prevalence of Vitamin A Deficiency in Populations at Risk 1995–2005*. WHO, Geneva, Switzerland.

Xu, X., Lin, H. and Fu, Z. (2004) Probe into the method of regional ecological risk assessment – a case study of wetland in the Yellow River Delta in China. *Journal of Environmental Management* 70(3), 253–262.

Yan, J., Kandianis, C. B., Harjes, C. E., Bai, L., Kim, E. H., *et al*. (2010) Rare genetic variation at *Zea mays* crtRB1 increases β-carotene in maize grain. *Nature Genetics* 42(4), 322–329.

Yan, W., Hunt, L., Sheng, Q. and Szlavnics, Z. (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science* 40, 597–605.

Zare, M., Choukan, R., Heravan, E.M., Bihamta, M.R. and Ordookhani, K. (2011) Gene action of some agronomic traits in corn (*Zea mays* L.) using diallel cross analysis. *African Journal of Agricultural Research* 6(3), 693–703.

Zhang, K., Tian, J., Zhao, L., and Wang, S. (2008) Mapping QTLs with epistatic effects and QTL × environment interactions for plant height using a doubled haploid population in cultivated wheat. *Journal of Genetics and Genomics* 35(2), 119–127.

Zhang, Y. and Kang, M.S. (2003) DIALLEL-SAS: A program for Griffing's diallel methods. In: Kang, M.S. (ed) *Handbook of Formulas and Software for Plant Geneticists and Breeders*. Haworth Press Inc., New York, pp.1–19.

Zhang, Y., Kang, M.S., and Lamkey, K.R. (2005). DIALLEL-SAS05: A comprehensive program for Griffing's and Gardner-Eberhart analyses. *Agronomy Journal* 97, 1097–1106.

18 Developments in Genomics Relative to Abiotic Stress-tolerance Breeding in Maize During the Past Decade

M.T. Labuschagne*
University of the Free State, Bloemfontein, South Africa

Introduction

Maize (*Zea mays* L.) is a staple food crop to many countries in Africa and South America. Maize grain yield in sub-Saharan Africa remains low at an average of 1.8 t ha^{-1} for 2011–2013, compared to 3.1 t ha^{-1} in Mexico (http://faostat3.fao.org). These low yields are mainly attributable to drought and low fertilizer use, as well as to other biotic and abiotic stresses (Shiferaw *et al.*, 2011). In Asia, more than 80% of maize production is under rainfed conditions, making it very vulnerable to climate variability, but very little breeding emphasis has been placed on heat and drought tolerance (Vivek *et al.*, 2017).

Maize yield losses attributable to drought are estimated to be about 25%. There is a large gap between potential yield and actual yield, which can be reduced by 20–25% by breeding for drought tolerance (Edmeades, 2013). Breeding for genotypes that are adapted to areas prone to water stress is critical because of the impact of climate changes and limited water resources. Newly developed hybrids should be able to withstand drought stress during the growing season but should not have a yield penalty under optimum conditions (Abdulmalik *et al.*, 2017). CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) and their partners have used conventional breeding through the Water Efficient Maize for Africa (WEMA) project to release more than 200 drought-tolerant hybrids and open-pollinated varieties (OPVs) for sub-Saharan Africa (DTMA, 2015).

Most genomics-related research in maize abiotic stress breeding in the past has been focused on drought stress (Semagn *et al.*, 2013; Xue *et al.*, 2013; Cooper *et al.*, 2014; Thirunavukkarasu *et al.*, 2014; 2017; Zhang *et al.*, 2015; Beyene *et al.*, 2015; 2016; Zaidi *et al.*, 2016; Wang *et al.*, 2016; Abdulmalik *et al.*, 2017; Shikha *et al.*, 2017; Vivek *et al.*, 2017; Dias *et al.*, 2018; Nepolean *et al.*, 2018). Heat stress-tolerance breeding through genome-wide association studies is also underway in the Heat Stress Tolerant Maize for Asia (HTMA) project. This programme is a public–private alliance aimed at the development and deployment of heat-resilient maize hybrids for resource-poor people and smallholder farmers (https://www.cimmyt.org/project-profile/heat-tolerant-maize-for-asia).

The aim of this chapter is to give an overview of the developments in the application of genomics in maize abiotic stress-tolerance breeding in recent years. As most of the reported research so far has been on drought tolerance, with a few reports on heat tolerance, this was the main focus of this review.

* Email: labuscm@ufs.ac.za

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

Maize Phenotyping

Until recently, all abiotic-stress breeding was done using conventional breeding, based on plant phenotype and selection. Grain yield and factors contributing to yield were used for direct selection. Identification of, and selection for, traits that are highly heritable, which can be used in high throughput phenotyping and are highly positively correlated with yield traits, are key to improving drought tolerance (Maazou *et al.*, 2016). Precision phenotyping of these traits can be very difficult, but new high-throughout techniques have been developed, such as visible light imaging, thermal imaging of shoots and leaves, near-infrared imaging of plants or plant parts and shoot 3D imaging, as well as remotely controlled unmanned aerial devices (Nepolean *et al.*, 2018).

Genotypes being developed for drought tolerance are tested under water-managed environments; for example, optimal and water-stress conditions. Effective phenotypic screening is often time-consuming and expensive (Dias *et al.*, 2018) and most drought-tolerance-related traits are highly influenced by the environment and are determined by many genes with small effects (Zhang *et al.*, 2015). Phenotyping is currently one of the major bottlenecks in genomics-related breeding (Varshney *et al.*, 2018).

The Advent of Genomics in Plant Breeding

The total number of polymorphic markers identified in plant breeding increased significantly in the 1980s because of the development of molecular-marker systems. Initially, molecular markers were used together with phenotypic data for marker-assisted selection (MAS). However, this has proven unsuccessful for complex traits across multiple environments because of the presence of quantitative trait loci (QTL) × environment interaction (QEI) and different genetic backgrounds (Crossa *et al.*, 2017). Grain yield under drought conditions is very complex and influenced by genetic background and the environment; therefore, the use of single QTL with marker-assisted backcrossing (MABC) has also not been very effective. QTL associated with drought tolerance often only explain a small amount of phenotypic variation and are often genetic background-specific (Semagn *et al.*, 2013).

With the availability of next-generation sequencing technologies, genotyping has moved to high-throughput, single-nucleotide polymorphism (SNP)-based systems. SNPs are the markers of choice for genomic studies, as they are abundant and can be detected with high-throughput methods. SNPs have been used extensively in the discovery of QTL. High-density SNPs are necessary for high-resolution fingerprinting, genome-wide association study (GWAS) and genomic selection (GS). Low-to-medium-density SNPs are necessary for genetic diversity analysis, QTL/trait mapping, MAS, marker-assisted recurrent selection (MARS) and candidate gene-based selections (Nepolean *et al.*, 2018).

Marker-assisted Recurrent Selection

MARS is used in breeding to accumulate favourable genes from several genomic regions within a population (Abdulmalik *et al.*, 2017). MARS can be applied by selecting among F_2 recombinants or backcross progenies. In 2007, African maize breeding programmes started enhancing drought tolerance with MARS. MARS across ten populations in sub-Saharan Africa led to a yield gain of 0.051 t ha^{-1} (Beyenne *et al.*, 2016).

Abdulmalik *et al.* (2017) crossed maize inbred lines with resistance to *Striga* and drought tolerance. The F_1 was selfed to obtain the F_2, which was planted in the field to generate $F_{2:3}$ lines, which were crossed to an inbred tester from an opposite heterotic group. The test crosses were evaluated under well-watered and drought conditions. Marker effects of the $F_{2:3}$ lines were calculated with best linear unbiased prediction (BLUP) across lines, and multiplied with the marker score of each line, resulting in genomic-estimated breeding values (GEBVs). Three cycles of recurrent selection were applied (C_1, C_2 and C_3). They selected a total of 233 markers that were uniform and homozygous in the parents and polymorphic between parents, which were used for genotypic selection in the MARS populations. Phenotyping of the field trials was done in different locations and seasons. The mean frequency of favourable marker alleles for grain yield increased by 9% between C_0 and C_3. MARS was able to accumulate favourable alleles linked to desired QTL in the breeding population and decrease the frequency of unfavourable alleles.

Genome-wide Association Studies

Biparental populations, such as F_2-derived populations, recombinant inbred lines (RILs), near-isogenic lines (NILs) and others, which follow the principle of linkage, have been used previously for QTL mapping (Yan *et al.*, 2011; Thirunavukkarasu *et al.*, 2013). These populations provide information on two alleles per locus and are not very useful for finding marker–trait association, because of low recombination rates and limited genetic variation. This approach is time-consuming, as selection cycles are long, and marker–QTL associations are often absent for genes with minor effects (Xu *et al.*, 2012).

Association mapping started to replace linkage mapping in the early 2000s, which allowed the detection of marker–trait associations in populations not derived from two parents (Leng *et al.*, 2017). High-density SNP markers have opened the way for GWAS, which can overcome constraints posed by conventional linkage mapping, as a complementary strategy to study complex traits. GWAS combines high-throughput phenotypic and genotypic data to provide insights into genetic architecture of complex traits in maize (Yan *et al.*, 2011). It is a powerful tool for QTL mapping, as a broad range of genetic resources can be evaluated for marker–trait association, as there is no limit on the number of markers available (Leng *et al.*, 2017). Multi-parent populations, such as GWAS panels, rest on the principle of linkage disequilibrium (LD). This approach shortens the time to develop populations and allows the testing of more alleles (Thirunavukkarasu *et al.*, 2013). GWAS provides breeders with numerous marker–trait associations that can be exploited directly in breeding programmes, as they can be applied to a wide germplasm base, as long as high LD is maintained between the causal gene and significant markers in breeding material (Huang and Han, 2014).

Genome-wide association mapping in maize

Drought tolerance is conditioned by polygenes with additive effects from numerous chromosomal regions. To improve drought tolerance, these regions should be transferred to the target germplasm (Abdulmalik *et al.*, 2017). GWAS has been used increasingly because of the availability of reference genome sequences and high-density automated genotyping platforms (Xiao *et al.*, 2017). Association analysis is done with the abundant phenotypic variation in maize and high density of polymorphisms at a DNA level. With the help of high-density genotyping platforms and genotyping by sequencing techniques, millions of marker data points can be identified throughout the genome to use for effective GWAS (Pandey *et al.*, 2013; Suwarno *et al.*, 2014). Wang *et al.* (2016) used GWAS in maize to identify the drought-tolerance gene *ZMVPP1* and 42 candidate genes in seedlings exposed to drought stress.

Zaidi *et al.* (2016) used an association mapping panel of 396 diverse tropical maize lines to phenotype various structural and functional root traits under drought and well-watered conditions. This panel was genotyped with 955,690 SNPs and GWAS was done using 331,390 SNPs filtered from the entire SNP set. A total of 50 SNPs for root functional traits, 67 for root structural traits, and 28 for grain yield and shoot biomass were identified under well-watered and drought-stress conditions. The most SNPs were found on chromosome 5 (nine SNPs), 3 and 7 (eight SNPs each) for root functional traits, and on chromosome 1 and 9 had nine SNPs each followed by chromosomes 2 and 7 (eight SNPs each) for root structural traits under well-watered and drought-stress conditions. Regions on chromosomes 3 and 8 were reported to have association with drought tolerance and adaptation, as they had five and six meta QTL for grain yield and anthesis-silking interval, respectively.

The principle of genomics-assisted breeding (GAB) is to identify the best haplotypes and specific genomic regions to facilitate crop improvement (Zhong *et al.*, 2009; Xu *et al.*, 2012; Leng *et al.*, 2017). The results from association analysis can be used to predict the best haplotypes for one or multiple genes for optimum expression of the target trait (a haplotype is a set of polymorphisms, which can be alleles or SNPs, on the same chromosome, that tend to be inherited together) (Almeida *et al.*, 2013). Finding new haplotypes and selecting superior haplotypes can advance plant breeding (Spindel *et al.*, 2016). The first maize haplotype was constructed by Gore *et al.* (2009). This study identified and genotyped millions of sequence polymorphisms in 27 diverse maize inbred lines. They showed that the

genome was characterized by highly divergent haplotypes. GWAS was used to identify multiple haplotypes that were significantly associated with heat tolerance, including nine significant haplotype blocks (about 200 kb) for grain yield, which explained 4–12% of phenotypic variation individually (https://www.cimmyt.org/project-profile/heat-tolerant-maize-for-asia/). Calus *et al.* (2008) reported that haplotypes made up of ten markers gave the best estimation accuracy for breeding values.

Genomic selection

There has been rapid development in whole-genome sequencing and marker-based technology during the past decade, which has allowed the use of high-density SNP markers to analyse the whole genome at a very low cost (Leng *et al.*, 2017). Next-generation sequencing techniques have allowed the determining of as many loci as possible across the entire genome, with nucleotide precision (Varshney *et al.*, 2014, 2018). Appropriate GS methods can accurately predict performance even for untested genotypes, which can lead to progress in breeding programmes. It reduces the number of field trials and phenotyping costs. GS advantages are more when traits are polygenic and difficult to test, and where many environments are needed for testing (Krchov and Bernardo, 2015). GS predicts GEBVs by analysing traits and high-density marker scores with an artificially created population on a whole-genome level (Meuwissen *et al.*, 2001; Heffner *et al.*, 2009; Crossa *et al.*, 2017).

The GS model assumes that all marker loci in the genome contribute to trait expression, either positively or negatively, which means that small-effect loci will also be included in the model (Meeuwissen *et al.*, 2001; Guo *et al.*, 2012). The cumulative effect of SNPs is called GEBV, and this determines the expression of the trait. GS consists of two components, the prediction of GEBVs and utilization of GEBVs in the breeding programme. GEBVs can be predicted with GS models using genome-wide SNPs and comprehensive phenotypic data (Nepolean *et al.*, 2018; Wang *et al.*, 2018). GS incorporates all available marker information into a model to predict genetic value of progeny for selection, to facilitate the prediction of phenotype from genotype. So the effects of all markers are estimated simultaneously from a training population that was phenotyped and genotyped (Meuwissen *et al.*, 2001; Lorenz, 2013). GS reduces selection time by half per cycle when compared with phenotypic selection for almost all traits in maize (Lorenzana and Bernardo, 2009).

Every trait locus can be in LD with at least one marker locus in the whole target population; therefore, QTL–marker loci associations become obsolete in GS. QTL information generated in one generation is also often not valid in another, but in GS, the QTL detection step is eliminated (Meuwissen *et al.*, 2001). GS is superior to MARS for improving complex traits, as it effectively avoids issues related to the number of QTL related to a trait (Leng *et al.*, 2017).

The training population can consist of related individuals that are phenotyped and genotyped. The breeding population usually consists of progeny from the training population, or a variety related to the training population, and is only genotyped, not phenotyped. GS depends on the degree of genetic similarity between the breeding and training populations and the LD between marker and trait loci. Phenotyping is crucial to GS to build accurate statistical models (Desta and Ortiz, 2014).

In maize breeding, the breeder can test cross 50% of available lines, and evaluate them in first stage multi-location trials, and then use phenotypic data to predict the other 50% by GS. GS significantly reduces cost of test cross formation and evaluation at each stage in multi-location trials. A second cycle of selection using the training population from the previous cycle can be used, for example, to predict new doubled-haploid (DH) lines, thereby excluding test cross formation and stage 1 multi-location trials. Based on GS, the best lines can go directly to the second stage of multi-location trials. Parental average is from pedigree information, which allows compiling of a matrix between the individuals. Genetic gain and selection response per unit time is improved. GS is also useful for higher heritable traits and for predicting additive effects in early generations (such as $F_{2:3}$ lines) to get a rapid, short interval selection cycle. GS is highly useful in hybrid breeding, because maize hybrid genotype can be inferred from the inbred parents (Kadam *et al.*, 2016). Therefore, GS can partially

replace field testing if it is effectively integrated into the breeding process (Heffner *et al.*, 2010). Genotyping can be done in the off-season to save time. Zhang *et al.* (2017) were the first to study a multi-parental population, using 18 elite tropical maize inbred lines, which were intercrossed twice, and then selfed to form the training population. One thousand ear-to-row C_0 families were genotyped with dense genotyping by sequencing (GBS) markers, and test crosses were phenotyped at four locations to develop genomic prediction models. From C_1 to C_4, realized grain yield increase was 0.225 t ha^{-1} per cycle, with two rapid cycles per year.

Genomic selection in maize stress-tolerance breeding

In the past years, GS research in crops focused on developing and testing different statistical prediction models that can be applied in breeding to predict diverse breeding panels for different traits in different environments (Crossa *et al.*, 2010). In contrast to QTL and association mapping, GS uses all molecular markers for genomic prediction of the performance of possible plants for selection and can increase genetic gains, as it can accurately predict how untested genotypes will perform depending on markers distributed throughout the genome (Dias *et al.*, 2018). Different cross-validation designs were used to attempt prediction of performance of untested individuals and environments (Crossa *et al.*, 2010; Heslot *et al.*, 2013). Several statistical models were examined for GS in diverse maize panels by CIMMYT, using a random cross-validation scheme that mimics prediction of unobserved phenotype based on markers and pedigrees (Crossa *et al.*, 2010).

Through WEMA, CIMMYT has developed more than 34 biparental populations since 2009. Beyene *et al.* (2015) used test crosses derived from these populations under drought stress and well-watered conditions, followed by advancing of populations using MARS or GS. Eight of the 34 biparental populations were improved using GS, which were used for studies. They used BLUP as a predictive model (Hayes *et al.*, 2009) and a genomic relationship matrix according to VanRaden (2008). They reported a yield increase of 0.086 t ha^{-1} per cycle under drought stress without significant changes in maturity and plant height. With the exception of one population, all populations showed a consistent yield increase across selection cycles. The yield gain was two- to four-fold higher with GS compared with conventional selection techniques under drought stress. GS also offered considerable time saving over conventional breeding, as three cycles of GS could be completed in 1 year. This proved that GS was more effective under drought stress in tropical maize for improving yield than pedigree breeding. This approach will be useful to improve stress resilience in maize (Beyene *et al.*, 2015).

Vivek *et al.* (2017) started with drought-tolerance breeding in Asia by evaluating test cross performance of $F_{2:3}$ families (the training set), which was then used to select plants for recombination in a later generation (Cycle 1 test set). They generated cycle 1 (C1) from test cross data of $F_{2:3}$ families. Cycle 2 was derived from C1 through recombination based on selected plant characteristics under optimal conditions (pedigree breeding). Cycle 2 (test cross GS) was derived from a combination of C1 plants with high GEBVs for good test cross performance under drought and optimal conditions. C1 was the source for both these populations, which could then be compared for conventional and molecular breeding approaches. Both C2 populations showed improvement over the F_2. Only the C2-GS recombined plants carried genomic regions for drought tolerance. This population was better than that created with conventional breeding in the absence of the target stress. The use of GEBVs allowed the selection of better phenotypes in the absence of the target stress and led to rapid gains in drought tolerance. With conventional breeding, another four seasons would have been necessary (Vivek *et al.*, 2017).

Statistical models in genomic selection

The accuracy of GEBVs is determined by model performance, sample size, relatedness, marker density, gene effects, heritability and genetic architecture, and extent and distribution of LD between markers and QTL (Desta and Ortiz, 2014). In GS statistical models, complications occur because of the number of markers being more than the population size and high correlation

between markers. This can be corrected by using penalized regression, variable selection and reduction of dimensionality. Statistical models can also be used to assess genomic-enabled prediction complexities and high-density marker platforms with genotype by environment (G×E) interactions (Crossa *et al.*, 2017).

Different methods are used to determine breeding values from GS models. Parametric methods include Ridge Regression (RR), least absolute shrinkage and selector operator (LASSO), Elasticnet, Bayes A, Bayes B; semi-parametric methods include Reproducing Kernel Hilbert Space (RKHS) and non-parametric methods include Random Forest (RF) (Wang *et al.*, 2018).

Bayesian LASSO or BL, and RR models are the best for predicting GEBVs (Desta and Ortiz, 2014). Using different means, various GS models capture different aspects of the association of genotype with phenotype. The performance of different models depends on the genetic architecture of underlying specific traits. Various models can also be combined to improve efficiency (Wang *et al.*, 2018).

Accuracy in predicting traits that are affected by a large number of loci depends on the size and genetic diversity in the training population and its relationship with the testing population (whether they are relatives). The heritability of the traits is important, where traits with low heritability and small marker effects are suitable for GS (Daetwyler *et al.*, 2010). Depending on the trait, a plateau is reached in GS accuracy with an increase in the population size (Lorenz *et al.*, 2012).

Most of the current genomic prediction studies only apply a single environment model, assuming that environments are correlated (Guo *et al.*, 2013). GS models should take G×E interaction into account using statistical-genetic models exploiting multi-trait, multi-environment variance-covariance and genetic correlations between environments, and between traits and environments, simultaneously (Heslot *et al.*, 2014; Oakey *et al.*, 2016; Crossa *et al.*, 2017). Maize single-cross hybrids were highly accurately predicted using models including G×E interaction (Technow *et al.*, 2014; Kadam *et al.*, 2016). Zhang *et al.* (2015) also showed the advantage of modelling for G×E interaction to predict untested genotypes. An extension of genomic BLUP (GBLUP), incorporating G×E interaction has improved accuracy of predicting unobserved cultivars in environments and led to substantial increases in prediction accuracy of unobserved individuals in different environments (Heslot *et al.*, 2014; Cuevas *et al.*, 2016).

Jarquin *et al.* (2014) developed models that incorporate random structures of high dimensional environment and marker information. Using their model, Zhang *et al.* (2015), in a study using 19 biparental maize populations evaluated under several drought and well-watered environments, and genotyped with low density and GBS SNPs, found that mean breeding values derived from G×E interaction models were higher than corresponding values from non-G×E interaction models for all cross-validation, marker densities and trait–environment combinations, especially for complex traits. Across the populations, the difference between G×E interaction and non-G×E interaction models was consistent under drought and well-watered conditions. For the less complex traits, such as days to anthesis and plant height, the G×E interaction model was not superior.

Burgueño *et al.* (2012) were the first to use both marker- and pedigree-based GBLUP models to assess G×E interaction for genomic-enabled prediction. Kernel-based methods, such as RKHS, have led to good genomic predictions in plants. These GS prediction models can be used to develop heat- and drought-tolerant plants by exploiting positive G×E interactions (Crossa *et al.*, 2017).

In breeding programmes, the focus is on additive and total genetics effects, although non-additive effects (dominance and epistasis) are also important to understand the genetic architecture of target traits, and to devise optimal breeding strategies. Estimating additive effects and corresponding variance components is difficult, and they rely on appropriate mating designs and many observations (Dias *et al.*, 2018). Studies have shown that molecular-based relationship matrices can improve orthogonality and predictability of additive and non-additive effects (Muñoz *et al.*, 2014; Nazarian and Gezan, 2016). Dominance effects must also be included in models for maize, where heterosis is present, such as in single-cross hybrids (De Almeida Filho *et al.*, 2016; dos Santos *et al.*, 2016).

Dias *et al.* (2018) showed in tropical maize germplasm in Brazil that a high level of predictive

accuracy is possible in untested single-cross hybrids for drought-related traits by including G×E interaction, additive and dominance effects together in a multi-environmental trial (MET) model that incorporates genomic relationship matrices (Oakey *et al.*, 2016). Dias *et al.* (2018) extended the GBLUP model to account for additive and dominance effects in the context of MET data using factor analytic structures. The presence of high G×E interaction for most drought tolerance-related traits highlights the importance of using models that can deal with MET data and can model G×E interaction. Beyene *et al.* (2015) followed a similar approach to show the advantage of GS over phenotypic selection to increase genetic gain in maize drought-tolerance breeding. Dias *et al.* (2018) suggested the use of GBLUP models that account for additive and dominance effects routinely in MET analysis for prediction of performance of untested hybrids for drought tolerance in maize breeding programmes.

Genotyping by Sequencing

GBS platforms are next-generation sequencing-based, and can be applied to crops without prior genomic knowledge, or data on ploidy level or genome size. GBS generally describes all platforms that use a sequencing approach for genotyping, such as Elshire GBS, diversity array technology sequencing (DArT-seq), sequence-based genotyping (SBG) and restriction enzyme site comparative analysis (RESCAN), of which the first two are the most widely used platforms in crop genomics. GBS reduces genome complexity by using restriction enzymes to cleave the DNA, coupled with DNA-barcoded adapters, PCR and sequencing (Rasheed *et al.*, 2017). It generates high-density genome-wide markers through tagging randomly shared unique, short, DNA sequences (barcodes) (Zhang *et al.*, 2015). The advent of GBS has increased the availability of molecular markers from about a hundred to thousands of SNPs distributed evenly through the genome (Poland *et al.*, 2012). Therefore, the confidence interval of QTL was reduced, which allowed the development of genetic maps with high resolution and precise QTL mapping.

The sequenced portion of the genome is highly consistent within a population, because restriction sites are generally conserved across species. This makes GBS a powerful tool for implementing GWAS, genomic diversity studies, genetic linkage analysis, molecular marker discovery and GS in plant breeding programmes (Rasheed *et al.*, 2017).

Co-dominant markers, such as SNP, give a more accurate estimate of GEBV than dominant markers, such as DArT, because of their better LD detection power (Barabaschi *et al.*, 2016). By 2009, the correlation between true breeding value and GEBV was already reported to have increased to 0.85, even in traits with low heritability (Heffner *et al.*, 2009).

Zhang *et al.* (2015) reported that, compared to low density SNPs (about 200 markers), GBS improved prediction ability. Prediction of all target traits under stress conditions was lower than that under well-watered conditions, and multi-environment models incorporating G×E interaction were more accurate for complex traits, such as grain yield. Prediction accuracy of GBS was better than that of low-density SNPs for complex traits under drought and well-watered conditions.

Cerrudo *et al.* (2018) used GBS technology to detect QTL and develop GS models for yield and related traits in a biparental DH line. Full-sib recurrent selection was done under drought conditions for seven cycles to develop the drought-tolerant parent line. The other parent line had several good characteristics, such as disease resistance and agronomic traits. The doubled haploids were derived from this cross. The lines were test crossed to CML494 for phenotypic evaluation. The doubled haploids and the test crosses were evaluated under drought and optimal conditions at different locations for three seasons. Training and validation sets were created to assess prediction accuracy. A bin map with 191 bins was constructed using high-quality filtered GBS SNPs. Neighbouring SNPs with high similarity haplotype information were clustered into one bin. Each bin was treated as a single marker to construct the genetic map. A total of 48 significant QTL for nine traits were identified. None of the 39 QTL detected for secondary traits overlapped for hybrids or lines or across treatments. There was an association of QTL in some bins, and bin 1.02 was shown as important for genetic control of grain yield and vigour. Genomic prediction accuracy was defined as the average correlation value between phenotype and GEBVs.

Of 48 QTL found for grain yield and secondary traits, none was consistent in hybrids and lines, as lines and hybrids were poorly correlated and there was G×E interaction, epistasis and heterosis. They emphasized the need to phenotype the test crosses, not the lines, to find QTL to be used in hybrids. A single QTL for yield was consistent across treatments. Prediction accuracy for yield was better under well-watered than under drought conditions, and accuracy was better for secondary traits than for yield. Prediction accuracy for GS-MAS was higher than all variation explained by all QTL in QTL-MAS for grain yield and the secondary traits. Detected QTL in bins 1.02, 1.03 and 7.04 can be used for forward-breeding to enrich alleles for these traits in the breeding programme for line conversions (Cerrudo *et al.* 2018). GS-MAS can be used in more mature breeding programmes to also capture alleles with smaller additive effects (Cao *et al.*, 2017).

Trachsel *et al.* (2016) reported association of QTL for plant height and senescence. Prediction accuracy under well-watered conditions was better than under drought stress for yield and secondary traits. There was a positive correlation between GS prediction accuracy and trait heritability for well-watered hybrids. This was not the case under drought conditions.

Chip-based technologies, such as Diversity Array Technology (DArT), and SNP have laid a strong foundation for application of GS in crop breeding (Wang *et al.*, 2018). The use of GBS and SNP chips is important in trait-associated markers, which are then used for gene tagging and gene pyramiding. Automated chip-based platforms are useful for genome-wide association linkage analysis and genetic diversity analysis (Rasheed *et al.*, 2017).

Resequencing

Rapid progress in next-generation sequencing techniques has led to more sequenced species (Leng *et al.*, 2017). Reference genome sequences have been published for many crops. This provides a starting point for exploring the genome and provides information on genetic variation through partial or complete resequencing of different accessions. Resequencing leads to arrays of high-density SNPs, which allow whole-genome scans to identify haplotype blocks significantly related to quantitative trait variation. Then GWAS becomes attractive to map QTL, as all genetic resources can be scanned for marker–trait associations without any limits on markers. The high marker number also supports GS (Barabaschi *et al.*, 2016).

Whole-genome sequences of ten maize lines are currently available, including reference lines B73, W22 and Mo17 (www.maizegdb.org). Genome-wide SNPs can be used for identification of haplotypes and for genetic mapping (Nepolean *et al.*, 2018). Maize haplotype 1 (Hapmap1) was developed in 2009 (Gore *et al.*, 2009), Hapmap2 in 2012 (Chia *et al.*, 2012) and currently Hapmap3, which has a size of 3.83 million SNPs and InDels, and was identified in 1218 maize genotypes (Bukowski *et al.*, 2015).

Worldwide, there is an effort to understand the genetic basis of agronomic traits, providing catalogues of allele series for the most important loci, which will allow breeders to select the best allele combinations. The accumulation of QTL data will allow meta-analyses, which can provide consistent determinants of quantitative traits in crop breeding by exploiting and integrating datasets from different studies (Barabaschi *et al.*, 2016). The available low-cost markers can be introgressed into breeding material from landraces and wild accessions with limited linkage drag (Varshney *et al.*, 2014). By using germplasm collections and crop wild relatives, traits can be mapped that are climate change-relevant, using high-throughput genotyping and phenotyping platforms. This can assist in developing climate change-ready genotypes (Varshney *et al.*, 2018).

Genome Editing

Genome editing allows specific nucleotides in a genome to be changed, which can be used to generate homozygous mutants for multiple target genes in one generation, which is faster than even molecular breeding techniques (Varshney *et al.*, 2018). Genome editing can be used as an alternative to normal breeding processes through recombination and genetic transformation. The clustered regularly interspaced short palindromic repeats/CRISPR-associated protein-9 nuclease (CRISPR/Cas9) system is currently being used for precise genome editing (Miglani, 2017).

Editing on the gene promoter region can give rise to differential cis-transcription alleles, creating new quantitative variation in plant breeding (Leng *et al.*, 2017). Editing target genes could improve drought and other abiotic stress tolerance (Nepolean *et al.*, 2018).

Illumina developed a high-density Illumina® MaizeSNP50 Beadchip (Wu *et al.*, 2014). An array with lower density was later developed from the same platform with a 55,229 SNP density, which covers tropical and temperate maize germplasm (Xu *et al.*, 2017). These SNP chips were used for genetic characterization of maize inbreds (Thirunavukkarasu *et al.*, 2013) and in GWAS (Li *et al.*, 2013; Thirunavukkarasu *et al.*, 2014).

Conclusions

The ample plant genome information that is currently available has generated many next-generation sequencing-based approaches for allele mining and for identifying candidate genes in breeding programmes. High-throughput trait-associated markers, genotyping approaches and phenotyping platforms are facilitating genomics-assisted breeding. The research in maize so far has largely focused on drought tolerance, with some research on heat tolerance, but will probably in the future extend to breeding for tolerance to other important abiotic stress conditions. GS with high-throughput phenotyping will become routine in plant breeding programmes in future. GS can also have an application in forming gene pools and populations from gene bank accessions, which could be a rich source of new alleles, especially with the prospect of severe climate change effects. Identification of new haplotypes and selection of superior haplotypes can advance maize breeding for abiotic stress tolerance. Improved varieties with biotic and abiotic stress tolerance and good yield may result from new genetic variants, rare alleles and new haplotypes. Coupling of genome-wide haplotypes with GS is feasible to accelerate maize breeding, as they are specific and accurate. GS combined with GWAS has significant potential, whereby haplotypes can be identified with GWAS, and then used to identify promising lines with good GEBVs.

Currently, there are inconsistencies in genotyping platforms and germplasm used, and in statistical procedures. Genome-wide meta-analysis, supported by new statistical procedures, and availability of reference genomes are becoming powerful tools to integrate information that can reduce redundancy of information.

The progress in genomics in maize breeding, and especially drought-tolerance breeding, has been rapid during the past 10 years. Significant research has been done, especially in Africa, where maize is the major staple crop for millions of people, and where climate change is already having a significant effect on farming activities. Genomics-assisted breeding has the potential to speed up the breeding process to develop climate-resilient maize genotypes for production by small-scale farmers and communities, who rely on maize for a livelihood. The technology is consistently developing, and is already revolutionizing plant breeding, and the speed of development will probably increase in coming years.

References

Abdulmalik R.O., Menkir, A., Meseka, S.K., Unachukwu, N., Ado, S.G., *et al*. (2017) Genetic gains in grain yield of a maize population improved through marker assisted recurrent selection under stress and non-stress conditions in West Africa. *Frontiers in Plant Science* 22, 841. DOI: 10.3389/fpls.2017.00841.

Almeida, G.D., Makumbi, D., Magorokosho, C., Nair, S., Borém, A., *et al*. (2013) QTL mapping in three tropical maize populations reveals a set of constitutive and adaptive genomic regions for drought tolerance. *Theoretical and Applied Genetics* 126, 583–600. DOI: 10.1007/s00122-012-2003-7.

Barabaschi, D., Tondelli, A., Desiderio, F., Volante, A., Vaccino, P., *et al*. (2016) Next generation breeding. *Plant Science* 242, 3–13. DOI: 10.1016/j.plantsci.2015.07.010.

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., *et al*. (2015) Genetic gains in grain yield through genomic selection in eight biparental maize populations under drought stress. *Crop Science* 55, 154–163. DOI: 10.2135/cropsci2014.07.0460.

Beyene, Y., Semagn, K., Crossa, J., Mugo, S., Atlin, G.N., *et al*. (2016) Improving maize grain yield under drought stress and non-stress environments in sub-Saharan Africa using marker-assisted recurrent selection. *Crop Science* 56, 344–353. DOI: 10.2135/cropsci2015.02.0135.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., *et al*. (2015) Construction of the third generation *Zea mays* haplotype map. *GigaScience* 7, 1–12. DOI: 10.1093/gigascience/gix134.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J.L. (2012) Genomic prediction of breeding values when modelling genotype–environment interaction using pedigree and dense molecular markers. *Crop Science* 52, 707–719. DOI: 10.2135/cropsci2011.06.0299.

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W. and Veerkamp, R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553–561. DOI: 10.1534/genetics.107.080838.

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., *et al*. (2017) Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *Plant Genome* 10, 1–14. DOI: 10.3835/plantgenome2016.10.0099.

Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E.A., *et al*. (2018) Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Frontiers in Plant Sciences* 9, 366. DOI: 10.3389/fpls.2018.00366.

Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., *et al*. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics* 44, 803–807. DOI: 10.1038/ng.2313.

Cooper, M., Gho, C., Leafgren, R., Tang, T. and Messina, C. (2014). Breeding drought tolerant maize hybrids for the US-cornbelt: Discovery to product. *Journal of Experimental Botany* 65, 6191–6204. DOI: 10.1093/jxb/eru064.

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., *et al*. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. DOI: 10.1534/genetics.110.118521.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., *et al*. (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science* 22, 961–975. DOI: 10.1016/j.tplants.2017.08.011.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, Pérez-Rodríguez, *et al*. (2016) Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* 9, 1–20. DOI: 10.3835/plantgenome2016.03.0024.

Daetwyler, H.D., Pong-Wong, R., Villaneuva, B., and Woolliams, J.A. (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. DOI: 10.1534/genetics.110.116855.

De Almeida Filho J.E., Guimarães, J.F., Silva, F.F., de Resende, M.D., Muñoz, P., *et al*. (2016) The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity* 117, 33–41. DOI: 10.1038/hdy.2016.23.

Desta, Z.A. and Ortiz, R. (2014) Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science* 19, 592–601. DOI: 10.1016/j.tplants.2014.05.006.

Dias, K.O.D.G., Gezan. S.A., Guimarães. C.T., Nazarian. A., *et al*. (2018) Improving accuracies of genomic predictions for drought tolerance in maize by joint modelling of additive and dominance effects in multi-environment trials. *Heredity* 121, 24–27. DOI: 10.1038/s41437-018-0053-6.

dos Santos, J.P.R., Vasconcellos, R.C.dC., Pires, L.P.M., Balestre, M. and Von Pinho, R.G. (2016) Inclusion of dominance effects in the multivariate GBLUP model. *PLoS ONE* 11(4), e0152045. DOI: 10.1371/journal.pone.0152045.

DTMA (2015) A new generation of maize for Africa. Available at: http://dtma.cimmyt.org/index.php/publications/doc_view/196-a-newgeneration-of-maize-for-africa (accessed 4 March 2019).

Edmeades, G.O. (2013) *Progress in Achieving and Delivering Drought Tolerance in Maize – an Update*. International Service for the Acquisition of Agri-Biotech Applications, Ithaca, New York.

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., *et al*. (2009) A first-generation haplotype map of maize. *Science* 80, 1115–1117. DOI: 10.1126/science.1177837.

Guo, Z., Tucker, D.M., Lu, J., Kishore, V. and Gay, G. (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theoretical and Applied Genetics* 124, 261–275. Doi: 10.1007/s00122-011-1702–9.

Guo, Z., Tucker, D.M., Wang, D., Basten, C.J., Ersoz, E., *et al*. (2013) Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. *G3: Genes, Genomes, Genetics* 3, 263–272. DOI: 10.1534/g3.112.005066.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92, 433–43. DOI: 10.3168/jds.2008-1646.

Heffner, E.L., Lorenz, A.J. and Jannink, J.L. (2009) Genomic selection for crop improvement. *Crop Science* 49, 1–12. DOI: 10.1016/j.cj.2018.03.001.

Heffner, E. L., Lorenz, A. J., Jannink, J. L. and Sorrells, M.E. (2010) Plant breeding with genomic selection: Gain per unit time and cost. *Crop Science* 50, 1681–1690. DOI: 10.2135/cropsci2009.11.0662.

Heslot, N., Jannink, J.L. and Sorrells, M.E. (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Science* 53, 921–933. DOI: 10.2135/cropsci2012.07.0420.

Heslot, N., Akkdemir, D., Sorrels, M.E. and Jannink, J.L. (2014) Integrating environmental covariates and crop modelling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics* 127, 463–489. DOI: 10.1101/014100.

Huang, X. and Han, B. (2014) Natural variations and genome-wide association studies in crop plants. *Annual Review in Plant Biology* 65, 531–551.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., *et al*. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127, 595–607. DOI: 10.1146/annurev-arplant-050213-035715.

Kadam, D.C., Potts, S.M., Bohn, M.O., Lipka, A.E. and Lorenz, A.J. (2016) Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. *G3: Genes, Genomes, Genetics* 6, 3443–3453. DOI: 10.1534/g3.116.031286.

Krchov, L.M. and Bernardo, R. (2015) Relative efficiency of genome wide selection for testcross performance of double haploid lines in a maize breeding program. *Crop Science* 55, 2091–2099. DOI: 10.2135/cropsci2015.01.0064.

Leng, P., Lübberstedt, T. and Xu, M. (2017) Genomics-assisted breeding – a revolutionary strategy for crop improvement. *Journal of Integrative Agriculture* 16, 2674–2685. DOI: 10.1016/S2095-3119(17)61813-6.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., *et al*. (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature Genetics* 45, 43–50. DOI: 10.1038/ng.2484.

Lorenz, A.J. (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. *G3: Genes, Genomes, Genetics* 3, 481–491. DOI: 10.1534/g3.112.004911.

Lorenz, A.J., Smith, K.P. and Jannink, J. (2012) Potential and optimization of genomic selection for *Fusarium* head blight resistance in six-row barley. *Crop Science* 52, 1609–1621. DOI: 10.2135/cropsci2011.09.0503.

Lorenzana, R.E. and Bernardo, R. (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120, 151–161. DOI: 10.1007/s00122-009-1166.

Maazou, Q.S., Tu, J., Qiu, J. and Liu, Z. (2016) Breeding for drought tolerance in maize (*Zea mays* L.). *American Journal of Plant Sciences* 7, 1858–1870. DOI: 10.4236/ajps.2016.714172.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Miglani, G.S. (2017) Genome editing in crop improvement: Present scenario and future prospects. *Journal of Crop Improvement* 31(4), 453–559. DOI: 10.1080/15427528.2017.1333192.

Muñoz, P.R., Resende, J.R.M.F.R., Gezan, S.A., Resende, M.D.V., De Los Campos, G., *et al*. (2014) Unravelling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198, 1759–1768.

Nazarian, A. and Gezan, S.A. (2016) Integrating non-additive genomic relationship matrices into the study of genetic architecture of complex traits. *Journal of Heredity* 107, 153–162. DOI: 10.1534/genetics.114.171322.

Nepolean, T., Kaul, J., Mukri, G. and Mittal, S. (2018) Genomics-enabled next-generation breeding approaches for developing system-specific drought tolerant hybrids in maize. *Frontiers in Plant Science* 9, 361. DOI: 10.3389/fpls.2018.00361.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., *et al*. (2016) Genomic selection in multi-environment crop trials. *G3: Genes, Genomics*, *Genetics* 6, 1313–1326. DOI: 10.1534/g3.116.027524.

Pandey, N., Ranjan, A., Pant, P., Tripathi, R.K., Ateek, F., *et al*. (2013) CAMTA 1 regulates drought responses in *Arabidopsis thaliana*. *BMC Genomics* 14, 216. DOI: 10.1186/1471-2164-14-216.

Poland, J.A., Brown, P.J., Sorrells, M.E. and Jannink, J. (2012) Development of high-density genetic maps for barley and wheat using a novel two enzyme genotyping-by-sequencing approach. *PLoS ONE* 7, e32253. DOI: 10.1371/journal.pone.0032253.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., *et al*. (2017) Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. *Molecular Plant* 10, 1047–1064. DOI: 10.1016/j.molp.2017.06.008.

Semagn, K., Beyene, Y., Warburton, M.L., Tarekegne, A., Mugo, S., *et al*. (2013) Meta-analyses of QTL for grain yield and anthesis silking interval in 18 maize populations evaluated under water-stressed and well-watered environments. *BMC Genomics* 14, 313. DOI: 10.1186/1471-2164-14-313.

Shiferaw, B., Prasanna, B.M., Hellin, J. and Bänziger, M. (2011) Crops that feed the world. Past successes and future challenges to the role played by maize in global food security. *Food Security* 3, 307–327. DOI: 10.1007/s12571-011-0140-5.

Shikha, M., Kanika, A., Rao, A.R., Mallikarjuna, M.G., Gupt, H.S., *et al*. (2017) Genomic selection for drought tolerance using genome-wide SNPs in maize. *Frontiers in Plant Science* 8, 550. DOI: 10.3389/fpls.2017.00550.

Spindel, J.E., Begum, H., Akdemir, D., Collard, B., Redoña, E., *et al*. (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. DOI: 10.1038/hdy.2015.113.

Suwarno, W.B., Pixley, K.V., Palacios-Rojas, N., Kaeppler, S.M. and Babu, R. (2014) Formation of heterotic groups and understanding genetic effects in a provitamin A biofortified maize breeding program. *Crop Science* 54, 14–24. DOI: 10.2135/cropsci2013.02.0096.

Technow, F., Schrag, T.A., Schipprack, W., Bauer, E., Simianer, H., *et al*. (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197, 1343–1355. DOI: 10.1534/genetics.114.165860.

Thirunavukkarasu, N., Hossain, F., Kaliyugam, S., Swati, M., Kanika, A., *et al*. (2013) Unravelling the genetic architecture of subtropical maize (*Zea mays* L.) lines and their utility in breeding programs. *BMC Genomics* 14, 877. DOI: 10.1186/1471-2164-14-877.

Thirunavukkarasu, N., Hossain, F., Arora, K., Sharma, R., Shiriga, K., *et al*. (2014) Functional mechanisms of drought tolerance in subtropical maize (*Zea mays* L.) identified using genome-wide association mapping. *BMC Genomics* 15, 1182. DOI: 10.1186/1471-2164-15-1182.

Thirunavukkarasu, N., Sharma, R., Singh, N., Shiriga, K., Mohan, S., *et al*. (2017) Genomewide expression and functional interactions of genes under drought stress in maize. *International Journal of Genomics* 2017, 1–14. DOI: 10.1155/2017/2568706.

Trachsel, S., Sun, D., SanVicente, F.M., Zheng, H., Atlin, G.N., *et al*. (2016) Identification of QTL for early vigor and stay-green conferring tolerance to drought in two connected advanced backcross populations in tropical maize (*Zea mays* L.). *PLoS ONE* 11(3), e0149636. DOI: 10.1371/journal.pone.0149636.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91, 4414–4423. DOI: 10.3168/jds.2007-0980.

Varshney, R.K., Terauchi, R. and McCouch, S.R. (2014) Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biology* 12 (2014), e1001883. DOI: 10.1371/journal.pbio.1001883.

Varshney, R.K., Singh, V.K., Kumar, A., Powell, W. and Sorrells, M.E. (2018) Can genomics deliver climate-change ready crops? *Current Opinion in Plant Biology* 45, 205–211. DOI: 10.1016/j.pbi.2018.03.007.

Vivek, B.S., Krishna, G.K., Vengadessan, V., Babu, R., Zaidi, P.H., *et al*. (2017) Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize. *The Plant Genome* 10, 1–8. DOI: 10.3835/plantgenome2016.07.0070.

Wang, X., Wang, H., Liu, S., Ferjani, A., Li, J., *et al*. (2016) Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. *Nature Genetics* 8, 1233–1241. DOI: 10.1038/ng.3636.

Wang, X., Xu, Y., Hu, Z. and Zu, C. (2018) Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal* 6, 330–340. DOI: 10.1016/j.cj.2018.03.001.

Wu, X., Li, Y., Shi, Y., Song, Y., Wang, T., *et al*. (2014) Fine genetic characterization of elite maize germplasm using high-throughput SNP genotyping. *Theoretical and Applied Genetics* 127, 621–631. DOI: 10.1007/s00122-013-2246-y.

Xiao, Y., Liu, H., Wu, L., Warburton, M. and Yan, J. (2017) Genomewide association studies in maize: Praise and stargaze. *Molecular Plant* 10, 359–374. DOI: 10.1016/j.molp.2016.12.008.

Xu, C., Ren, Y., Jian, Y., Guo, Z., Zhang, Y., *et al*. (2017) Development of a maize 55K SNP array with improved genome coverage for molecular breeding. *Molecular Breeding* 37, 20. DOI: 10.1007/s11032-017-0622-z.

Xu, Y., Lu, Y., Xie, C., Yanli, L., Gao, S., *et al*. (2012) Whole-genome strategies for marker-assisted plant breeding. *Molecular Breeding* 29, 833. DOI: 10.1007/s11032-012-9699-6.

Xue, Y., Warburton, M.L., Sawkins, M., Zhang, X., Setter, T., *et al*. (2013) Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. *Theoretical and Applied Genetics* 126, 2587–2596. DOI: 10.1007/s00122-013-2158-x.

Yan, J., Warburton, M. and Crouch, J. (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. *Crop Science* 51, 433–449. DOI: 10.2135/cropsci2010.04.0233.

Zaidi, P.H., Seetharam, K., Krishna, G., Krishnamurthy, L., Gajanan, S., *et al*. (2016) Genomic regions associated with root traits under drought stress in tropical maize (*Zea mays* L.). *PLoS ONE* 10, 1371. DOI: 10.1371/journal.pone.0164340.

Zhang, X., Perez-Rodriguez, P., Semagn, K., Beyene, Y., Babu, R., *et al*. (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity* 114, 291–299. DOI: 10.1038/hdy.2014.99.

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., *et al*. (2017) Rapid cycling genomic selection in a multiparental tropical maize population. *G3: Genes, Genomes, Genetics* 7, 2315–2326. DOI: 10.3389/fpls.2017.01916

Zhong, S.Q., Dekkers, J.C.M., Fernando, R.L. and Jannink, J.L. (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* 182, 355–364. DOI: 10.1534/genetics.108.098277.

19 Exploiting Alien Genetic Variation for Germplasm Enhancement in *Brassica* Oilseeds

Mehak Gupta and S.S. Banga*
Punjab Agricultural University, Ludhiana, India

Introduction

Domesticated crops owe their existence to multiple rounds of unconscious selection by ancient farmers, which transformed prevalent wild species into edible and farmworthy landraces. Modern crop breeders then converted these landraces into qualitatively superior modern crop varieties. Ancient selection involved retaining seeds from a small number of relatively better-looking or better-performing plants for future sowings by proto-farmers. This gradually increased the frequency of plants with desirable traits, e.g. increase in number or size of seeds, loss of seed shattering, uniform maturity, modified plant architecture and transition from perennial to annual forms (Konishi *et al.*, 2006; Hua *et al.*, 2015). Unlike domestication, the next rounds of changes were propelled by the knowledge of genetics. Plant breeders purposefully started making controlled crosses between phenotypically superior individuals followed by artificial selection to combine the traits of interest and advancing only a limited number of desirable genotypes. Modern crop biotechnologies have further enhanced the selection efficiency and speeded up the breeding process. All these gains have, however, caused severe genetic bottlenecks, as evident from loss of allelic diversity at loci under selection. This is reflected in declining response to selection in most important crops. As in the past, the challenge of feeding growing and aspirational population in the face of climate change is a daunting task. As per the predicted models (The Global Risks Report, 2019), rising global temperatures, increasing sea levels and increased levels of atmospheric CO_2 are expected to cause changes in rainfall patterns, salinity, loss of existing species diversity, and arrival of new pests and pathogens.

Plant breeders require abundant genetic variation to meet these challenges. Wild and weedy crop species constitute rich reservoirs of genetic variation (Buckler *et al.*, 2001; Miller and Gross, 2011; Meyer *et al.*, 2012) and can potentially help in widening the genetic base of current crops. As per Harlan (1976), there are several examples in which genes from wild relatives stand between man and starvation. Success stories include the transfer of resistance against late blight (caused by *Phytophthora infestans* (Mont.) de Bary) of potato from the wild potato (*Solanum demissum* Lindl.) into cultivated potato, thus mitigating the effects of the Irish famine in 1945, and resistance against stem rust (caused by *Puccinia graminis* ssp. *graminis*) from the wild wheat *Aegilops tauschii* Coss. into cultivated wheat, which sustained the green revolution (Prescott-Allen and Prescott-Allen, 1986; Kilian *et al.*, 2010).

* Email: nppbg@pau.edu

© CAB International, 2020.

It is also known that the wild relatives, the so-called inferior species, are also a source of yield-related genes that are primarily targeted for breeding plant varieties. Recently, Celik *et al*. (2017) have helped in detecting 37 quantitative trait loci (QTL) for 11 fruit-quality traits in an inbred backcross line (IBL) population derived from the cross between wild species *Solanum pimpinellifolium* and an elite cultivar of tomato (*Solanum lycopersicum*), out of which desirable alleles for 16 QTL were contributed by wild species. In spite of low value of fruit weight, internal colour and stem scar in the wild relative (*S. pimpinellifolium*), it donated favourable alleles for these traits. By realizing the significance of crop wild relatives, plant breeders are increasingly exploring their use as a source of novel traits through 'pre-breeding' strategies to retune the genetic diversity of cultivated crops by restoring ancient genetic variation.

Family Brassicaceae

Brassicaceae was previously known as Cruciferae; it includes about 338 genera, of which *Brassica* is economically the most important. It includes six crop species that were domesticated as edible oilseeds, vegetables, spices and condiments for human consumption, and as forage crops for livestock feeding. This genus belongs to the subtribe Brassicinae, one among nine subtribes in the tribe Brassiceae (Gomez-Campo, 1999a; Prakash, 2010). U's triangle (Nagaharu, 1935) depicts the cytogenetic relationship among six main cultivated species of the genus *Brassica*, of which *Brassica nigra* L. Koch (2n = 16; BB), *Brassica oleracea* L. (2n = 18; CC) and *Brassica rapa* L. (2n = 20; AA) represent the three diploid species. Pairwise intercrossing among them gave rise to three allotetraploid species, *Brassica carinata* A. Braun (2n = 34; BBCC), *Brassica juncea* (L.) Czern. and Coss. (2n = 36; AABB), and *Brassica napus* L. (2n = 38; AACC). Among the diploid species, *B. rapa* and *B. oleracea* exhibit vast morphological diversity. Sub-species of *B. rapa* have been classified as rapifera (oilseed forms of Europe and Canada), oleifera (oilseed forms of Indian subcontinent) and leafy types (China and other South-east Asian countries). *Brassica nigra* or black mustard, also prevalent in Europe as a weed, is grown as a condiment crop. *Brassica oleracea* got diversified into many botanical groups and related crops during domestication. These include: var. *acephala,* var. *botrytis,* var. *capitata,* var. *gemmifera,* var. *gongylodes,* var. *italica* and var. *sabauda;* also known by the names of kale, cauliflower, cabbage, Brussels sprout, kohlrabi, broccoli and Savoy cabbage crops (Linnaeus, 1753; Lamarck, 1784; De Candolle, 1821). All these forms of *B. oleracea* are well-known vegetable crops worldwide. Among the amphidiploid species, *B. carinata*, known as Ethiopian mustard, resulted from the union of B genome (from *B. nigra*) and C genome (from *B. oleracea*). It has a natural distribution in the Abyssinian Plateau, where it is cultivated as a source of edible oil, spices, medicinals and vegetables. *Brassica juncea* (Indian mustard) is a key source of edible oil in the Indian subcontinent and East European countries. It is cultivated primarily as a vegetable crop (leaf mustard) in China, and as a hot mustard condiment in Europe, Canada and America. *Brassica napus* is a polyploid of recent origin. It orginated through multiple hybridization events between *B. rapa* and *B. oleracea*. It is extensively grown in Europe, Canada, China and Australia for edible oil. Though significant gains have been made in the past in the seed yields and quality (canola) modification, no progress has been made for resistance to biotic or abiotic stresses. No genotype of *Brassica* is recognized with apomictic traits worldwide. These goals can only be achieved through wise utilization of ancestral wild relatives, which are a repository of many valuable genes.

Brassica coenospecies

These comprise nine genera (*Diplotaxis, Brassica, Eruca, Erucastrum, Hirschfeldia, Coincya, Sinapis, Sinapidendron* and *Trachystoma*) from subtribe Brassicinae, along with two genera from subtribe Raphaninae (*Raphanus, Enarthocarpus*) and three genera from subtribe Moricandiinae (*Rytidocarpus, Moricandia, Pseuderucaria*) (Gomez-Campo, 1999a,b). These are more closely related to crop *Brassica* than other species of the family, as confirmed by a long series of research on the chloroplast DNA (cp-DNA) and restriction sites (Warwick and Black, 1991; Pradhan *et al*., 1992; Warwick *et al*., 1992; Warwick and Sauder, 2005). They have the capacity to readily exchange genetic material with crop *Brassica*. Wild relatives of *Brassica* are distributed across a

broad range, from the western Mediterranean area to the eastern end of the Sahara Desert in the north-west of India. These species have evolved in diverse ecological habitats, such as coastal dunes, slopes of coastal volcanos, stony pastures and arid to semi-arid regions across a large number of years (Tsunoda, 1980). Warwick *et al.* (2009) has published a guide of wild germpool of *Brassica* that provides useful information on their growth behaviour, chromosome number, geographical distribution and the traits of interest. Useful traits associated with various wild Brassicaceae species are described in Table 19.1. *Orychophragmus violaceus* (L.), a wild species, has a large number of primary branches, ranging between 13.4 and 14.9, more siliques, more seeds per silique (39.0–39.7) and bigger seed size (3.8–3.98 g per 1000 seeds) than other wild *Brassicaceae* species. Yields may go up to 2047.5–2085 kg ha^{-1} (Luo *et al.*, 1991). The seed oil also possesses high oleic acid (20.3%), linoleic acid, palmitic acid (14.3%) and low linolenic acids (4.8%). It has low (< 0.9%) erucic acid (Li *et al.*, 1995, 1996, 1998a,b, 2003, 2005a; Wu *et al.*, 1997; Ma *et al.*, 2006; Ma and Li, 2007; Xu *et al.*, 2007a,b; Zhao *et al.*, 2007, 2008; Ge *et al.*, 2009). *Hirschfeldia incana*, morphologically similar to black mustard, though an abnoxious weed, has immense potential for phytoremediating contaminated soils because of its ability to accumulate heavy metals in shoots and leaves. As the metal concentrations in the aerial parts of this species reflect the soil concentrations, it can serve as an indicator of heavy metal contamination in the soil (Gisbert *et al.*, 2008). *Crambe abyssinica* is another wild crucifer that has the capacity to tolerate and accumulate significant amounts of arsenic (As). Its seed oil is very useful for industrial purposes because of a high erucic acid content (55–60%), making it suitable for extracting waxes and developing base for paints, coatings, lubricants and many other products (Wang *et al.*, 2000). It is also drought and frost tolerant (Duke *et al.*, 1983; Zanetti *et al.*, 2013). *Eruca sativa*, rocket salad, is eaten raw in salads and it has excellent diuretic and antiscorbutic properties. It is also reported to be tolerant to drought and aphids (Tsunoda, 1980; Fahleson *et al.*, 1997). *Brassica villosa* is a source of beneficial glucosinolates, such as glucoiberin or glucoraphanin, that has anticarcinogenic properties (Faulkner *et al.*, 1998). *Camelina sativa* is drawing attention from plant breeders because of its exceptionally high (up to 45%) concentration of omega-3 fatty acids, which is often scarce in vegetable sources (Gugel and Falk, 2006). Its oil is also very rich in natural antioxidants, such as tocopherols, making it a highly stable oil, resistant to oxidation and rancidity (Sampath, 2009). *Capsella bursa-pastoris* is a traditional vegetable, medicinal, natural double-low (erucic acid, glucosinolates) species and found to be highly resistant to diseases, such as *Alternaria* blight, and *Sclerotinia* stem rot, and cold environment (Chen *et al.*, 2007a,b). Thus, the wild species of *Brassica* denote an extensive, divergent gene pool for the improvement of oilseeds *Brassica* whose genomes are highly plastic for genetic manipulation (Warwick and Black, 1991).

Wide Hybridization

Bypassing reproductive constraints is a key strategy to exploit alien germplasm for genetic enhancement of crop species. Systematic research on exploitation of wild germplasm through wide hybridization was first initiated by Mizushima during the 1950s, primarily to establish cross-genome homologies (Mizushima, 1950, 1968). Sageret (1826) was the first to produce an intergeneric hybrid between *Raphanus sativus* and *B. oleracea*. This was followed by the synthesis of Raphanobrassica by Karpechenko (1924) and Brassicaraphanus by Terasawa (1932). Since then, several alien species have been exploited by plant breeders by transferring genes for resistance, particularly to biotic and abiotic stresses as well as for improving seed quality traits through the production of cross-species hybrids, amphiploids and chromosome addition or direct introgression lines. Transfer of cytoplasm from *Brassica* wild relatives has proved crucial in the development of cytoplasmic-male sterility and fertility restoration (CMS-Rf) systems to commercialize hybrids in rapeseed-mustard crops.

Constraints and amendments for alien gene transfer

Efforts at trait introgression are generally limited by the incompatibility barriers like low

Table 19.1. Some of the examples of wild relatives of *Brassica* as potential sources of desirable traits.

Genus	Species	Potential trait (resistance/ improvement)	Reference
Arabidopsis	*thaliana*	Resistance to blackleg – *Leptosphaeria maculans* (*Phoma lingam*)	Brun and Tribodet (1995); Chen and Seguin-Swartz (1997, 1999)
		Clubroot – *Plasmodiophora brassicae*	Rehn *et al*. (2004)
		Flea beetles – *Phyllotreta cruciferae* and *Phyllotreta striolata*	Prakash and Bhat (2007)
		Acetolactate synthase (ALS) inhibiting herbicides (imidazolinone)	Roux *et al*. (2005a,b)
Arabis	*gunnisoniana*	Apomictic traits	Taskin *et al*. (2004)
Arabis	*holboellii*	Apomictic traits	Naumova *et al*. (2001)
Brassica	*cretica, incana, villosa*	Cabbage aphid – *Brevicoryne brassicae*	Kift *et al*. (2000)
		Cabbage white fly – *Aleyrodes proletella*	Ramsey and Ellis (1994)
		High erucic acid (>45–50%)	Yaniv *et al*. (1991, 1995)
Brassica	*fruticulosa*	Blackleg – *L. maculans*, black leaf spot – *Alternaria* spp.	Siemens (2002)
		Cabbage white fly – *Al. proletella*	Ramsey and Ellis (1994)
		Cabbage root fly or cabbage maggot – *Delia radicum*	Ellis *et al*. (1999)
		Cabbage aphid – *Br. brassicae*	Cole (1994); Ellis and Farrell (1995)
		Mustard aphid – *Li. erysimi*	Kumar *et al*. (2011); Atri *et al*. (2012)
		Sclerotinia stem rot – *Sclerotinia sclerotiorum*	Garg *et al*. (2010)
Brassica	*insularis*	Blackleg – *L. maculans*	Mithen *et al*. (1987); Mithen and Magrath (1992)
		Cabbage white fly – *Al. proletella*	Ramsey and Ellis (1994)
Brassica	*elongata*	Blackleg – *L. maculans*, black leaf spot – *Alternaria* spp.	Siemens (2002)
		High linoleic and linolenic acids	Velasco *et al*. (1998)
Brassica	*macrocarpa, hilarionis*	Resistance to pod shattering	Mithen and Herron (1991)
Brassica	*spinescens*	Black leaf spot – *Alternaria* spp.	Agnihotri *et al*. (1991)
		Cabbage aphid – *Br. brassicae*	Cole (1994); Ellis and Farrell (1995)
Brassica	*atlantica*	Blackleg – *L. maculans*	Mithen *et al*. (1987); Mithen and Magrath (1992)
Brassica	*tournefortii*	Cabbage seedpod weevil – *Ceutorhynchus obstrictus*	Ulmer and Dosdall (2006); Carcamo *et al*. (2007)
		Drought tolerance	Salisbury (1989); Prakash and Bhat (2007)
		Acetolactate synthase (ALS) inhibiting herbicides (sulfonylurea)	Adkins *et al*. (1997); Boutsalis *et al*. (1999)
		Resistance to pod shattering	Salisbury (1989)
Brassica	*maurorum*	Black leaf spot – *Alternaria* spp., white rust – *Alb. candida*	Chrungu *et al*. (1999)

Table 19.1. Continued.

Genus	Species	Potential trait (resistance/ improvement)	Reference
Brassica	*oxyrrhina, amplexicaulis*	High photosynthetic rates	Uprety *et al*. (1995)
Brassica	*rupestris*	High erucic acid (>45–50%)	Velasco *et al*. (1998)
Brassica	*souliei*	Black leaf spot – *Alternaria* spp.	Siemens (2002)
Barbarea	*vulgaris*	Cold tolerance	Laroche *et al*. (1992)
		Diamond-back moth – *Plutella xylostella*, European flea beetle – *Phyllotreta nemorum*, cabbage butterfly – *Pieris* spp.	Renwick (2002); Lu *et al*. (2004)
Coincya	*monensis*	Blackleg – *L. maculans*	Siemens (2002); Winter *et al*. (1999, 2003)
Crambe	*hispanica*	High erucic acid (>45–50%)	Yaniv *et al*. (1991, 1995); Prakash and Bhat (2007)
		Flea beetles – *Ph. cruciferae* and *Ph. striolata*	Soroka *et al*. (2003)
Crambe	*maritima*	Salt tolerance	Ashraf and Noor (1993); Ashraf (1994)
Crambe	*abyssinica*	Diamond-back moth – *Plu. xylostella*	Kmec *et al*. (1998)
		Flea beetles (*Ph. cruciferae* and *Ph. striolata*)	Anderson *et al*. (1992); Henderson *et al*. (2004)
		Proteinaceous seed meal	Carlson and Tookey (1983)
		High erucic acid (>45–50%)	Yaniv *et al*. (1991, 1995); Prakash and Bhat (2007)
		Arsenic tolerance	Paulose *et al*. (2007)
		Drought tolerance	Zanetti *et al*. (2013)
		Frost tolerance	Duke (1983)
Camelina	*microcarpa*	Acetolactate synthase (ALS) inhibiting herbicides (sulfonylurea)	Hanson *et al*. (2004)
Camelina	*sativa*	Black leaf spot – *Alternaria* spp.	Conn *et al*. (1988); Westman *et al*. (1999); Siemens (2002); Pedras and Adio (2008)
		Blackleg – *L. maculans*	Siemens (2002); Li *et al*. (2005b)
		Flea beetles – *Ph. cruciferae* and *Ph. striolata*	Soroka *et al*. (2003); Henderson *et al*. (2004)
		Mustard sawfly – *Athalia proxima*	Singh and Sachan (1997)
		High tocopherols and omega-3 fatty acids	Angelini *et al*. (1997); Zubr and Matthaus (2002); Gehringer *et al*. (2006); Gugel and Falk (2006)
		Drought resistance	Vollmann *et al*. (2005)
Capsella	*bursa-pastoris*	Clubroot – *Pl. brassicae*	Siemens (2002)
		Sclerotinia stem rot – *Sc. sclerotiorum*	Chen *et al*. (2007a,b)
		Black leaf spot – *Alternaria* spp.	Conn *et al*. (1988); Westman and Dickson (1998); Siemens (2002)
		Flea beetles – *Ph. cruciferae* and *Ph. striolata*	Prakash and Bhat (2007)

Continued

Table 19.1. Continued.

Genus	Species	Potential trait (resistance/ improvement)	Reference
		Photosystem II inhibiting herbicides (simazine)	Stanek and Lipecki (1991); Heap (2009)
		Double low fatty acids	Chen *et al*. (2007a,b)
Diplotaxis	*catholica*	Black leaf spot – *Alternaria* spp.	Prakash and Bhat (2007)
		High photosynthetic rates	Uprety *et al*. (1995)
Diplotaxis	*viminea*	High photosynthetic rates	Uprety *et al*. (1995)
		High tocopherols (vitamin E)	Goffman *et al*. (1999)
Diplotaxis	*acris, harra*	Drought tolerance	Boaz *et al*. (1990); Prakash and Bhat (2007)
Diplotaxis	*erucoides*	Black leaf spot – *Alternaria* spp.	Siemens (2002); Klewer *et al*. (2003)
Diplotaxis	*tenuifolia*	Black leaf spot – *Alternaria* spp.	Siemens (2002); Klewer *et al*. (2003)
		Blackleg – *L. maculans*	Chen and Seguin-Swartz (1997, 1999)
		C_3–C_4 intermediate	Apel *et al*. (1997); Bang *et al*. (2003); Ueno *et al*. (2003)
		Acetolactate synthase (ALS) inhibiting herbicides (sulfonylurea)	Heap (2009)
Diplotaxis	*muralis*	Blackleg – *L. maculans*	Chen and Seguin-Swartz (1997, 1999)
Erucastrum	*laevigatum*	High photosynthetic rates	Uprety *et al*. (1995)
Erucastrum	*cardaminoides*	High erucic acid (>45–50%)	Prakash and Bhat (2007)
		Sclerotinia stem rot – *Sc. sclerotiorum*	Garg *et al*. (2010)
Erucastrum	*gallicum*	Sclerotinia stem rot – *Sc. sclerotiorum*	Lefol *et al*. (1996)
Eruca	*sativa*	Root-knot nematode – *Meloidogyne* spp.	Dallavalle *et al*. (2005)
		Proteinaceous seed meal	Fagbenro (2004)
Enarthrocarpus	*lyratus*	High photosynthetic rates	Uprety *et al*. (1995)
Enarthrocarpus	*strangulatus*	Drought tolerance	Boaz *et al*. (1990)
Eruca	*vesicaria*	Downy mildew – *Peronospora parasitica*	Singh and Kolte (1999)
		Blackleg – *L. maculans*	Tewari *et al*. (1996); Siemens (2002)
		High erucic acid (>45–50%)	Yaniv *et al*. (1991, 1995)
		Black leaf spot – *Alternaria* spp.	Conn and Tewari (1986)
		Sclerotinia stem rot – *Sc. sclerotiorum*	Guan *et al*. (2004)
		White rust – *Alb. candida*	Bansal *et al*. (1997)
		Mustard aphid – *Li. erysimi* (Kalt.)	Rana *et al*. (1995); Chander and Bakhetia (1998)
		Cabbage aphid – *Br. brassicae*	Singh *et al*. (1994)
		Industrial oil value	Yaniv *et al*. (1998); Warwick *et al*. (2007)
		Salt tolerance	Ashraf and Noor (1993); Ashraf (1994)
		Drought tolerance	Prakash and Bhat (2007)
Eruca	*pinnatifia*	Blackleg – *L. maculans*	Tewari *et al*. (1996); Siemens (2002)

Table 19.1. Continued.

Genus	Species	Potential trait (resistance/ improvement)	Reference
Hirschfeldia	*incana*	Blackleg – *L. maculans*	Siemens (2002)
		Resistance to pod shattering	Salisbury (1989)
Lepidium	*perfoliatum*	Negative allelopathic effects	Aminidehaghi *et al*. (2006)
Lepidium	*sativum*	High linolenic acid	Prakash and Bhat (2007)
		Heavy metal tolerance and hyperaccumulation	Robinson *et al*. (2003)
		Medicinal properties	Mathews *et al*. (1993); Gokavi *et al*. (2004)
Lesquerella	*fendleri*	Salt tolerance	Dierig *et al*. (2001, 2004)
		High hydroxy fatty acids (lesquerolic acid)	Angelini *et al*. (1997)
		Proteinaceous seed meal	Wu and Hojilla-Evangelista (2005)
Lesquerella	*grandiflora*	Industrial oil value	Marvin *et al*. (2000)
Moricandia	*arvensis, nitens, sinaica*	C_3–C_4 intermediate	Bauwe (1983); Apel *et al*. (1997); Yan *et al*. (1999)
Orychophragmus	*violaceus*	High oleic and linoleic acids	Wang *et al*. (1999)
		High photosynthetic rates	Wu *et al*. (2007)
		High yield potential	Luo *et al*. (1994); Wang *et al*. (1999)
Pseuderucaria	*clavata*	Drought tolerance	Boaz *et al*. (1990)
Raphanus	*raphanistrum*	Blackleg – *L. maculans*	Chen and Seguin-Swartz (1999)
		Diamond-back moth – *Plu. xylostella*	Lehtila and Strauss (1999)
		Photosystem II inhibiting herbicides (triazines)	Walsh *et al*. (2004, 2007); Friesen and Powles (2007)
		Acetolactate synthase (ALS) inhibiting herbicides (sulfonylurea)	Hashem *et al*. (2001); Tan and Medd (2002)
Raphanus	*raphanistrum* ssp. *maritimus*	Salt tolerance	Inan *et al*. (2004)
Raphanus	*sativus*	White rust – *Alb. candida*	Williams and Pound (1963); Kolte *et al*. (1991)
		Black leaf spot – *Alternaria* spp., blackleg – *L. maculans*	Siemens (2002)
		Root-knot nematode – *Meloidogyne* spp.	Dallavalle *et al*. (2005); Pattison *et al*. (2006)
		Beet cyst nematode – *Heterodera schachtii*	Lelivelt and Krens (1992); Voss *et al*. (2000)
Sinapis	*alba*	Black leaf spot – *Alternaria* spp.	Sharma and Singh (1992); Siemens (2002)
		Blackleg – *L. maculans*	Gugel and Seguin-Swartz (1997)
		Turnip mosaic virus	Mamula *et al*. (1997)
		Flea beetles – *Ph. cruciferae* and *Ph. striolata*	Lamb (1980); Henderson *et al*. (2004)
		Cabbage aphid – *Br. brassicae*	Thompson (1963)
		Cabbage root fly or cabbage maggot – *D. radicum*	Jyoti *et al*. (2001)
		Cabbage seedpod weevil – *C. obstrictus*	Ulmer and Dosdall (2006); Carcamo *et al*. (2007)

Continued

Table 19.1. Continued.

Genus	Species	Potential trait (resistance/ improvement)	Reference
		Beet cyst nematode – *H. schachtii*	Lelivelt *et al*. (1993)
		Root-knot nematode – *Meloidogyne* spp.	Pattison *et al*. (2006)
		High erucic acid (>45–50%)	Yaniv *et al*. (1994, 1995)
Sinapis	*arvensis*	Turnip mosaic virus	Mamula *et al*. (1997)
		Blackleg – *L. maculans*, black leaf spot – *Alternaria* spp.	Siemens (2002); Winter *et al*. (1999, 2003)
		Dicamba/2,4D herbicide resistance	Warwick *et al*. (2000); Yajima *et al*. (2004); Jugulam *et al*. (2005); Mithila and Hall (2007)
		Triazine, metribuzin and acetolactate synthase inhibiting herbicides (sulfonylurea)	Ali *et al*. (1986); Heap (2009)
		High erucic acid (>45–50%)	Daun *et al*. (2003)
		High photosynthetic rates	Uprety *et al*. (1995)
Sinapidendron	*angustifolia*	High erucic acid (>45–50%)	Prakash and Bhat (2007)
Trachystoma	*ballii*	High photosynthetic rates	Uprety *et al*. (1995)
Thlaspi	*arvense*	Cold tolerance	Laroche *et al*. (1992); Sharma *et al*. (2007)
		Blackleg – *L. maculans*	Pedras *et al*. (2003)
		Flea beetles – *Ph. cruciferae* and *Ph. striolata*	Gavloski *et al*. (2000)
		Acetolactate synthase inhibiting herbicides (imidazolinone)	Beckie *et al*. (2007)
Thlaspi	*caerulescens*	Heavy metal tolerance and hyperaccumulation (Zn, Cd, Ni, Pb)	Pollard and Baker (1996, 1997); Guan *et al*. (2008)
Thlaspi	*montanum*	Heavy metal tolerance and hyperaccumulation (Zn, Ni)	Boyd and Martens (1998)

crossing success, F_1 sterility, hybrid breakdown, reduced chromosome pairing and linkage drag. Standardization of tissue culture techniques, such as embryo rescue and somatic hybridization, has helped in surmounting these barriers and a large number of interspecific/intergeneric hybrids have been produced. In spite of this, there are only a few examples that demonstrate introgression of defined traits from wild aliens into stable karyotypes of crop *Brassica*. This is primarily attributable to general lack of pairing between wild and crop genomes. Identification of a locus controlling homoeologous recombination and its suppression in *Brassica* (like the *Ph* locus in wheat) can promote the transfer of alien chromatin into cultivated *Brassica*. Another major challenge is the difficulty in detecting or isolating genes in the wild or unadapted germplasm and their introgression into crop *Brassica* without cotransfer of associated but undesirable gene complexes. The availability of whole-genome or transcriptome sequences at low cost now allows the development of species-specific markers and information on differentially expressed genes. This knowledge is very helpful in selecting which genetic factors or components of quantitative trait variation to introduce from a wild or alien gene pool into an elite cultivar. In the era of molecular breeding, targeted introgression is an eminently achievable objective for obviating the constraints of linkage drag, and even pyramiding genes from multiple wild species.

Sexual incongruity

Wide hybridization via controlled pollination is a prerequisite for introgressing alien genetic variation for crop improvement. However, independent evolution of different species either in the same, overlapping or different geographical ranges has led to development of a number of barriers termed as reproductive barriers that prevent the flow of alien chromatin into cultivated species (Sharma, 1995). The more phylogenetically diverse the species, the earlier will be the action of barriers. Even if a hybrid embryo forms, its development may be impeded by multiple barriers in a later stage. These barriers serve as isolating mechanisms established as consequences of evolutionary divergence of species to maintain their genetic integrity (Mayr, 1947). Stebbins (1958) classified reproductive barriers into pre-fertilization barriers and post-fertilization barriers, depending upon the stage of their occurrence during the course of hybridization.

PRE-FERTILIZATION BARRIERS These prevent mating and fertilization and may manifest in the absence of pollen recognition or germination on stigma at the time of pollination or anomalous growth of pollen tube in the style or lack of micropylar penetration of the female parent. One can possibly attempt to overcome these barriers by sampling broad genetic variation of the parental species in a very large number of hybrid combinations under a range of environmental conditions. This implies that crossability depends upon both genetic and environmental factors (Hermsen, 1984). This may be attributable to the presence of specific crossability genes with variable expression levels in response to diverse environmental conditions. For example, two genotypes of *B. rapa*, 'Shogoin-kabu' and 'Chiifu', show differential crossing ability to *R. sativus*. 'Shogoin-kabu' produces enough seeds, whereas 'Chiifu' is not able to set any seeds in crosses with *R. sativus*. Tonosaki *et al.* (2013) reported three QTL, qBrHFA-1, qBrHFA-2 and qBrHFA-3, that control embryo abortion. In another study (Li *et al.*, 1995), production of hybrids was dependent upon genetic composition of parental species. In a cross between *B. napus* and a wild species *O. violaceus,* only three *B. napus* genotypes gave rise to F_1s out of seven tried, whereas four other *B. napus* genotypes were sexually incompatible with *O. violaceus.*

Recent investigations have shown that pollen and pistil of respective crossed species undergo a complex series of cellular and molecular interactions during wide hybridization. This facilitates the pollen germination or growth of pollen-tubes from self-pollen or plants belonging to the same species and discourage invaders and less desirable pollen from another species (Bedinger *et al.*, 2017). This behaviour is similar to self-incompatibility (SI) mechanism found in diploid species of *Brassica*, controlled by a single, multiallelic *S* locus (Murfett *et al.*, 1996; Dresselhaus *et al.*, 2011). The only difference is the SI mechanism prevents self-mating in genotypes by rejecting their own pollen through precise genetic crosstalk between *S* allele-specific determinants of the male (*S*-locus protein 11/*S*-locus-cysteine-rich protein [SP11/ SCR]) and the female (*S*-locus receptor kinase [SRK]) haplotypes in the stigma to enforce outcrossing with other genotypes from the same species (Watanabe *et al.*, 2003). Earlier, Kerhaos *et al.* (1983) detected the callose deposition in intergeneric crosses when stigma of *B. napus* was pollinated by *Sinapsis arvensis* and prevented pollen hydration and germination. Similar kind of callose formation has been observed in stigmatic cells, which prevents penetration of pollen tube from self-pollen through style in self-incompatible *Brassica* species (Dumas and Knox, 1983). Another remarkable feature based on the number of interspecific or intergeneric crossability surveys is that crossibility is satisfactory only in one direction and more successful when the wild species is used as a female parent. In an investigation, pollen germination and pollen tube growth were normal in crosses when a wild species *Enarthocarpus lyratus* was used as female parent and cultivated Brassicas (*B. rapa, B. nigra, B. juncea, B. napus, B. carinata*) as male; in reciprocal crosses, however, when *E. lyratus* was the pollen parent, no pollen tube was seen in the style (Gundimeda *et al.*, 1992). Reciprocal differences for pollen germination and growth have been reported between *Diplotaxis siifolia* and crop Brassicas. It was normal on the stigma of *Di. siifolia*; however, in the reciprocal cross, the pollen tube of *Di. siifolia* failed to enter the stigma (Batra *et al.*, 1990; Ahuja *et al.*, 2003). In contrast, hybrids of *O. violaceus* with the *Brassica* species as maternal parent were all obtained without any trouble using embryo culture (Li *et al.*, 1995, 1998a,b; Li and

Heneen, 1999; Hua *et al.*, 2006). Nevertheless, when *O. violaceus* was used as the maternal parent, no hybrids were produced in spite of extensive attempts. This phenomenon of directional preference, called unilateral incompatibility, is commonly noticed in crosses between self-compatible and self-incompatible species where pollen from self-compatible species is often rejected by pistils of self-incompatible species, whereas the reciprocal crosses are fruitful (Murase *et al.*, 2004; Fujimoto *et al.*, 2006). Pollen from *B. oleracea* is accepted on stigmas of self-compatible lines of *B. rapa*, which lack functional SP11 and SRK, whereas it is rejected on pistils of self-incompatible lines of *B. rapa*. Cultivated tomato (*S. lycopersicum*) is self-compatible because of loss of the function mutation of *S* locus determinant male and female gametes (Kondo *et al.*, 2002). It is likely that the molecular mechanism controlling self-incompatibility and interspecific unilateral incompatibility may be linked or controlled by common genetic factors (Hiscock and Dickinson, 1993). Transfer of functional *S* molecules from self-incompatible wild tomato (*Solanum peruvianum*) into self-compatible cultivated ones caused interspecific, unilateral incompatibility in cultivated tomato (Tovar-Mendez *et al.*, 2014). In contrast, the genetic analysis of the F_2 populations between an interspecific-incompatible line and a self-compatible cultivar of *B. rapa* could not confirm it (Udagawa *et al.*, 2010). A wide range of artificial techniques, such as bud-pollination method (pollinating stigmas of immature buds 2–3 days before anthesis), stump pollination (cutting off the stigmatic surface, which is the primary site of barrier) and mixed pollination using mentor pollen (mixture of compatible [irradiated] and incompatible pollen), have been used to circumvent pre-fertilization hurdles in distant crossing. Bing (1991) was able to produce *B. juncea* × *Si. arvensis* hybrids and reciprocal hybrids of *Si. arvensis* × *B. nigra* using bud pollination technique. Sarla (1988) described the use of irradiated mentor pollen to overcome interspecific incompatibility in the cross between *B. campestris* ssp. *japonica* and *B. oleracea* var. *botrytis*. Use of various phytohormones, such as gibberellic acid, kinetin, napthalene acetic acid and indole acetic acid, on pollinated ovaries, has also been reported to delay abscission of the style till effective fertilization (Shivanna and Johri, 1985). A class of attractants, LUREs (cysteine-rich polypeptides), is secreted by synergid cells of embryo sac. These help in pollen tube guidance in a species-specific manner and do not attract pollen tubes of different species. The *AtLURE1* gene in the synergid cells of transgenic *Torenia fournieri* permitted the pollen tube penetration of *Arabidopsis thaliana* through the embryo sac of *T. fournieri,* suggesting that the expression of *AtLURE1* is enough to eliminate interspecific barriers (Takeuchi and Higashiyama, 2012). An Arabidopsis gene, *FERONIA* encoding receptor like kinase, is localized to the filiform apparatus of synergid cells that receives a male ligand and activates the feedback signal cascade from the synergid cell to the pollen tube for releasing sperm cells in the embryo sac (Escobar-Restrepo *et al.*, 2007). Mutation in this gene led to the overgrowth of the pollen tube, which was unable to burst in the female gametophyte (Huck *et al.*, 2003; Rotman *et al.*, 2003). A similar overgrowth of the pollen tube was observed in interspecific crossing of *A. thaliana* with pollen from a related Brassicaceae species. Understanding of these interacting components of male and female gametophytes at the molecular level may help to disrupt genetic communication between the two species, thereby overcoming pre-fertilization barriers in cross combinations. In some cases, an *in vitro* pollination technique is also used to obviate pre-fertilization barriers (Zenkteler, 1991, 2000). Kameya and Hinata (1970) produced hybrids between *B. chinensis* and *B. pikenensis* through pollination of excised, cultured ovules. More recently, Sosnowska and Cegielska-Taras (2014) were able to develop hybrids between *B. oleracea* and *B. rapa* using *in vitro* pollination of opened ovaries of *B. oleracea.*

POST-FERTILIZATION BARRIERS These include embryo and endosperm degeneration, abnormal growth and non-viability of hybrids, and their sterility. The primary cause of abortion of hybrid embryos is the shortage of nutrient supply to the zygote because of the failure of endosperm development after interspecific/intergeneric crossing (Van Tuyl *et al.*, 1991). Endosperm abortion is common in crosses between species with varied ploidy levels as compared to species having the same chromosome number or ploidy level. This is because of genome imbalance created by ploidy differences between the parents in the endosperm, resulting in underproliferation or overproliferation

of endosperm (Badger, 1988). Johnston *et al.* (1980) projected the endosperm balance number theory to decipher the basis for normal seed production. As per this theory, a proper ratio of maternal and paternal genetic material, i.e. 2:1, is critical in the endosperm for its normal development. Kinoshita (2007) has suggested that the endosperm abnormality in the F_1 hybrid may be caused by epigenetic misregulation of imprinting genes, expressed at varied levels in parental species. Extension of the syncytial phase, delay of cellularization of the endosperm (overproliferation) and arrest of embryo development (Chaudhury *et al.*, 1997; Grossniklaus *et al.*, 1998; Kiyosue *et al.*, 1999) have been reported in the mutants of endosperm imprinting genes, such as *MEDEA (MEA), FERTILIZATION INDEPENDENT SEED2 (FIS2)* and encoding components of the polycomb repressive complex 2 (PRC2). Some PRC2 components may also be associated with the post-zygotic barrier in the interspecific hybrids of Arabidopsis (Burkart-Waco *et al.*, 2013) or *Brassica* crops (Tonosaki *et al.*, 2013). Selecting higher ploidy species as the female parent can be useful for obtaining a balanced ratio of parental genomes in the endosperm and to moderate the misexpression of PRC2 genes (Johnston and Hanneman, 1982; Bushell *et al.*, 2003) to overcome endosperm lethality in wide crosses.

In vitro culture of hybrid embryos or embryo rescue is now a standard practice to raise rare hybrids (Zenkteler, 1990). For this, the embryo is dissected from an ovary or ovule and subjected to *in vitro* culture, obviating the need for the endosperm. This permits the maturation of the hybrid embryos. Brassicaceae species are very amenable to tissue culture techniques and respond well to embryo rescue. Embryo rescue can be done in one step (ovary culture/ovule culture/embryo culture) or in sequential steps (ovary culture, followed by ovule culture/ovule culture, followed by embryo culture or ovary culture, followed by ovule and embryo culture), depending upon the stage of abortion of the hybrid embryo. Ovary culture or ovule culture techniques are effective in wide crosses where the embryo abortion starts at a very initial phase of development. The embryo is cultured if embryo abortion occurs at an advanced stage of development (heart stage or later). Here, embryos are excised and positioned directly on to the culture medium. The sequential culture technique, which involves successive culturing of ovaries, ovules and embryos, seemed better for producing wide hybrids than single ovary/ovule/embryo culture (Vyas *et al.*, 1995). However, Mohanty *et al.* (2009) reported higher efficiency of ovule culture (16%) in obtaining intergeneric hybrids between *Erucastrum cardaminoides* and *B. oleracea* var. *alboglabra* in comparison to ovary culture (5.8%) and sequential culture (8.5%). Intergeneric hybrids from the cross *Sinapsis alba* × *B. carinata* were realized using sequential ovary–ovule culture (Sridevi and Sarla, 2005). Banga *et al.* (2003a) used sequential culture techniques to synthesize intergeneric hybrids between the wild crucifer *Diplotaxis catholica* and *B. rapa*/*B. juncea*. Intergeneric hybrids, *Diplotaxis erucoides* × *B. rapa* and *Brassica maurorum* × *B. rapa*, were developed using the sequential ovary–ovule culture by Garg *et al.* (2007). Almost 211 sexual hybrids, involving 45 wild species belonging to genera *Brassica, Sinapis, Diplotaxis, Moricandia, Eruca, Erucastrum, Enarthocarpus, Hirschfeldia, Capsella, Coincya, Crambe, Isatis and Orychophragmus* with crop *Brassica* (*B. juncea*, *B. napus*, *B. rapa*, *B. carinata* and *R. sativus*) have been synthesized using conventional crossing, aiding with various embryo-rescue techniques (Harberd and McArthur, 1980; Kaneko *et al.*, 2009).

Although crossing obstacles are circumvented in several cross combinations of *Brassica* and wild species through various artificial and embryo-rescue techniques, still a large number of alien species, distantly related to cultivated species, are underutilized for *Brassica* crop improvement because of sexual incongruity. In such cases, somatic hybridization, also known as parasexual hybridization, by the fusion of isolated protoplasts (from leaves, hypocotyls, cotyledons, roots, stems and microspores) of parental species, is a method of choice, which allows incorporation of desirable alleles from wild species into cultivated ones (Table 19.2). This method possesses additional merits over sexual hybridization in obtaining variation for cytoplasmically encoded traits (Glimelius, 1999; Christey, 2004; Li *et al.*, 2005a). The approach also offers added benefits in comparison with transgenic approaches by allowing the transfer of uncloned, multiple genes and generating products that are not subject to the same legal regulations as transgenic lines (Grosser and Gmitter, 2011). Protoplast fusion not only allows production of interspecific/intergeneric

Table 19.2. Traits targeted for introgressions through somatic hybridization between crop *Brassica* and wild aliens.

Participating genomes		Targeted traits	Reference
Between cultivated species			
B. oleracea (2n = 18)	*B. rapa* (2n = 20)	Synthesis of genetically diverse *B. napus*	Sundberg and Glimelius (1986)
		Synthesis of *B. napus* with low linolenic acid content	Heath and Earle (1997)
		Heat tolerance	Hossain and Asahira (1992)
		Atrazine herbicide resistance	Christey *et al*. (1991)
		Bacterial soft rot resistance	Ren *et al.* (2000)
	B. nigra (2n = 16), *B. juncea* (2n = 36), *B. carinata* (2n = 34)	Club root, blackleg, black spot and turnip mosaic virus resistance	Scholze *et al.* (2010) Wang *et al*. (2011a,b)
	B. nigra	Black rot resistance	
	B. napus (2n = 38)	Black rot resistance	Hansen and Earle (1995)
	Raphanus sativus (2n = 18)	Club root resistance	Hagimori *et al.* (1992)
B. juncea (2n = 36)	*B. oleracea* (2n = 18)	CMS source and *Verticillium dahliae* resistance	Lian *et al*. (2011)
B. napus (2n = 38)	*B. nigra* (2n = 16)	Blackleg and club root resistance	Sjodin and Glimelius (1989a)
	B. juncea (2n = 36), *B. carinata* (2n = 34)	Blackleg resistance	Sjodin and Glimelius (1989b)
	R. sativus (2n = 18)	CMS source	Sakai and Inamura (1990); Sundberg and Glimelius (1991)
		Beet cyst nematode resistance	Lelivelt and Krens (1992)
Intratribal			
B. rapa (2n = 20)	*Moricandia nitens* (2n = 28)	C_3–C_4 photosynthetic traits	Meng *et al.* (1999)
B. nigra (2n = 16)	*Sinapsis turgida* (2n = 18)	–	Toriyama *et al.* (1987b)
B. oleracea (2n = 18)	*Moricandia arvensis* (2n = 28)	C_3–C_4 photosynthetic traits	Toriyama *et al.* (1987a); Ishikawa *et al*. (2003)
	Mo. nitens (2n = 28)	C_3–C_4 photosynthetic traits	Yan *et al.* (1999)
	Si. turgida (2n = 18)	–	Toriyama *et al.* (1987b)
	Sinapsis alba (2n = 24)	Black spot resistance	Hansen and Earle (1994)
		Blackleg and black spot resistance	Ryschka *et al*. (1996)
		Club root resistance	Scholze *et al*. (2010)

Continued

Table 19.2. Continued.

Participating genomes		Targeted traits	Reference
B. juncea (2n = 36)	*Diplotaxis muralis* (2n = 42)	Drought tolerance	Chatterjee *et al.* (1988)
	Diplotaxis harra (2n = 26)	Drought tolerance	Begum *et al.* (1995)
	Diplotaxis catholica (2n = 18)	Black spot resistance, CMS source	Kirti *et al.* (1995a); Mohapatra *et al.* (1998)
	Eruca sativa (2n = 22)	Drought tolerance	Sikdar *et al.* (1990)
	Brassica spinescens (2n = 16)	High photosynthetic efficiency, white rust resistance and salt tolerance	Kirti *et al.* (1991)
	Mo. arvensis (2n = 28)	C_3–C_4 photosynthetic traits, white rust and black spot resistance	Kirti *et al.* (1992a)
	Trachystoma ballii (2n = 16)	Pod shattering and black spot resistance, CMS source	Kirti *et al.* (1992b, 1995b)
	Si. alba (2n = 24)	Black spot resistance	Gaikwad *et al.* (1996)
		Black spot resistance and heat stress tolerance	Kumari *et al*. (2018)
B. napus (2n = 38)	*Di. muralis* (2n = 42)	CMS source	McLellan *et al.* (1988)
	Di. harra (2n = 26)	CMS source	Klimaszewska and Keller (1988)
	Er. sativa (2n = 22)	High erucic acid content, aphid resistance, drought tolerance	Fahleson *et al.* (1988, 1997)
	Brassica tournefortii (2n = 20)	CMS source	Stiewe and Robbelen (1994)
		Blackleg resistance	Liu *et al.* (1995)
	Mo. arvensis (2n = 28)	C_3–C_4 photosynthetic traits	O'Neill *et al.* (1996)
	Mo. nitens (2n = 28)	C_3–C_4 photosynthetic traits	Meng *et al.* (1999)
	Sinapsis arvensis (2n = 18)	Blackleg disease resistance	Hu *et al*. (2002a)
	Si. alba (2n = 24)	Beet cyst nematode resistance	Lelivelt *et al*. (1993)
		Black spot resistance	Primard *et al.* (1988)
		High erucic acid	Wang *et al*. (2005a)
		Yellow seed colour	Li *et al.* (2012)
	Crambe abyssinica (2n = 90)	High erucic acid content	Wang *et al*. (2003, 2004)

Intertribal

Participating genomes		Tribe	Targeted traits	Reference
B. campestris syn. *B. rapa* (2n = 20)	*Arabidopsis thaliana* (2n = 10)	Sisymbrieae	Creation of novel species	Gleba and Hoffmann (1980); Hoffmann and Adachi (1981)

Continued

Table 19.2. Continued.

Intertribal

Participating genomes		Tribe	Targeted traits	Reference
	Barbarea stricta (2n = 16)	Arabideae	Cold tolerance	Oikarinen and Ryoppy (1992)
	Barbarea vulgaris (2n = 16)	Arabideae	Cold tolerance	Oikarinen and Ryoppy (1992)
	Isatis indigotica (2n = 14)	Lepidieae	Creation of genetic variation	Tu *et al.* (2008)
B. nigra (2n = 16)	*A. thaliana* (2n = 10)	Sisymbrieae	Creation of novel species	Siemens and Sacristan (1994, 1995)
B. oleracea (2n = 18)	*Armoracia rusticana* (2n = 32)	Arabideae	Club root resistance	Navratilova *et al*. (1997)
	Capsella bursa-pastoris (2n = 32)	Lepidieae	Flea beetle and black spot resistance	Sigareva and Earle (1997b, 1999b)
	Camelina sativa (2n = 40)	Sisymbrieae	Black spot resistance	Sigareva and Earle (1997a); Sigareva and Earle (1999a); Hansen (1997, 1998)
	Ba. vulgaris (2n = 16)	Arabideae	Clubroot, blackleg and black spot resistance	Ryschka *et al.* (1999)
	Matthiola incana (2n = 14)	Hesperideae	Clubroot, blackleg and black spot resistance	Ryschka *et al.* (1999)
B. juncea (2n = 36)	*Thlaspi caerulescens* (2n = 14)	Lepidieae	Toxic metal-resistant traits	Dushenkov *et al.* (2002)
	A. thaliana (2n = 10)	Sisymbrieae	Kanamycin (npt II) and phosphinothricin (bar) resistance	Ovcharenko *et al.* (2004)
B. carinata (2n = 34)	Ca. sativa (2n = 40)	Sisymbrieae	Black spot resistance	Narasimhulu *et al.* (1994)
B. napus (2n = 38)	*A. thaliana* (2n = 10)	Sisymbrieae	Acetolactate-synthase inhibiting herbicide resistance	Bauer-Weston *et al.* (1993)
			Black leg resistance	Forsberg *et al.* (1994, 1998a,b)
			Creation of novel species	Yamagishi *et al.* (2002)
	Thlaspi perfoliatum (2n = 42)	Lepidieae	High nervonic acid content	Fahleson *et al.* (1994a)
	Th. caerulescens (2n = 14)	Lepidieae	Zn and Cd tolerance	Brewer *et al.* (1999)
	Ba. vulgaris (2n = 16)	Arabideae	Cold tolerance	Fahleson *et al.* (1994b)
	Lunaria annua (2n = 28)	Lunarieae	High nervonic acid content	Craig and Millam (1995)
	Descurainia sophia (2n = 28)	–	High linolenic acid content, cold tolerance	Guan *et al.* (2007)
	Orychophragmus violaceus (2n = 24)	Orychophragmus	High linoleic and palmitic acid content	Hu *et al*. (2002b)
			Phosphinothricin resistance	Sakhno *et al.* (2007)

Continued

Table 19.2. Continued.

Intertribal				
Participating genomes		Tribe	Targeted traits	Reference
			Creation of genetic variation	Zhao *et al*. (2008)
	I. indigotica (2n = 14)	Lepidieae	Creation of genetic variation, secondary metabolites	Du *et al*. (2009)
	Lesquerella fendleri (2n = 12)	Drabeae	High lesquerolic acid content	Schroder-Pontoppidan *et al.* (1999)
			Transplastome transfer	Nitovskaya *et al.* (2006)
O. violaceus (2n = 24)	*Le. fendleri* (2n = 12)	Drabeae	Transplastome transfer	Ovcharenko *et al.* (2011)

hybrids but also helps produce a number of intertribal hybrids. Somatic hybridization based on protoplast fusion may be performed by symmetric fusion, asymmetric fusion or microfusion of protoplasts. In symmetric fusions, both the participating protoplasts contribute their complete genetic material, whereas in asymmetric fusions, the chromosomes of the donor protoplast are first fragmented by γ-, X- or UV-irradiation (Hall *et al*., 1992), or ultracentrifugation and then resulting fragments are integrated into the acceptor genome (Forsberg *et al*., 1998a,b). Hybrids from the asymmetric fusions are reported to be more successful than those from symmetric fusions in viability and regeneration ability because of smaller gene conflict between the partaking genomes and reduced number of backcrosses required to recover the recipient genome. For example, symmetric fusion between *B. napus* and *Lesquerella fendleri* yields sterile hybrids, whereas asymmetric somatic hybrids between the same parents were fertile and could set seed (Skarzhinskaya *et al*., 1996). Similarly, symmetric hybrids between *O. violaceus* and *B. napus* are self-sterile, whereas asymmetric hybrids are self-fertile (Hu *et al*., 2002b). Resulting advanced hybrid progenies from asymmetric cell fusions had higher levels of palmitic and linoleic acids, as well as reduced levels of erucic acid, in comparison to *B. napus*. In many experiments where symmetric fusions were attempted, target traits were detected in somatic hybrid progenies but not successfully incorporated in the euploid chromosome complement of cultivated species because of plant survival and sterility issues due to the large phylogenetic distance between the participating genomes. *Raphanus sativus* and *Si. alba* have been used for transferring resistance against beet cyst nematode to *B. napus* through protoplast fusion (Lelivelt and Krens, 1992; Lelivelt *et al*., 1993). Some of the somatic hybrids had a level of resistance as high as the resistant donor parent. However, no backcross and self progeny were produced because of high levels of sterility in hybrids. For transfer of resistance against Alternaria blight, somatic hybrids were produced between *B. carinata* and *Ca. sativa* (Narasimhulu *et al*., 1994). It was impossible to establish hybrid plants, as they failed to produce roots. Hansen (1998) reported the establishment of somatic hybrids obtained from protoplast fusion of rapid cycling *B. oleracea* and *Ca. sativa* after employing various tactics to help root induction, but plants failed to survive in the soil after 1 month out of *in vitro* culture. Somatic hybridization has been very successful in generating cytoplasmic male sterile sources and their corresponding fertility restorers by simultaneously allowing recombination between mitochondrial and nuclear genomes of wild and cultivated species.

ELIMINATION OF ALIEN CHROMATIN Following wide hybridization, complete or partial elimination of chromosomes from wild species has been reported. It may lead to haploid embryos and plants of one parent (Kasha and Kao, 1970) or a partial hybrid with a haploid complement from one parent and some chromosomes/segments of the other parent (Riera-Lizarazu *et al*., 1996). Tu *et al*. (2009) have reported partial hybrids in the progenies of an intertribal cross, *B. rapa* × *Isatis indigotica*. These partial hybrids carried a varying number of complete or partial chromosomes from the female parent. In another study, hybrids between *B. carinata*

and *O. violaceus* showed high pollen fertility and were mixoploids, with 2n chromosome number ranging from 17 to 35. F_1 plants closely resembled *B. carinata* in morphological attributes. Genomic *in situ* hybridization (GISH) analysis indicated the presence of the intact *B. carinata* genome in 2n = 34 or 35 F_1 plants and complete elimination of the *O. violaceus* genome. However, amplified fragment-length polymorphism (AFLP) analysis showed some bands specific for *O. violaceus*, suggesting some introgression from *O. violaceus* (Hua *et al.*, 2006). Such uniparental chromosome elimination may be caused by asynchronous cell cycles (Gupta, 1969), absence of synchrony in nucleoprotein synthesis, resulting in loss of the most retarded chromosomes (Bennett *et al.*, 1976, Laurie and Bennett, 1989), spatial separation of genomes during the interphase (Linde-Laursen and von Bothmer, 1999) and metaphase (Schwarzacher-Robinson *et al.*, 1987), formation of multipolar spindles (Subrahmanyam and Kasha, 1973), parent-specific inactivation of centromeres (Finch, 1983; Kim *et al.*, 2002; Jin *et al.*, 2004; Mochida *et al.*, 2004) and breakdown of alien chromosomes by host-specific nuclease activity (Davies, 1974).

HYBRID STERILITY Wide hybrids display male and female sterility caused by reduced chromosome pairing because of the occurrence of only a single copy of homologous chromosomes (Heslop-Harrison, 1999). This primarily leads to univalents, a small proportion of bivalents and rarely high-order interactions. Bivalents are mostly rod shape monochiasmatic and rarely multichiasmatic rings. Meiotic irregularities, such as laggards and anaphase bridges, may cause male and/or female sterilities (Stebbins, 1966). This sterility of F_1 hybrids restricts their role in genetic introgression. Normal meiosis and fecundity can be restored through somatic chromosome doubling by the use of cell cycle-arresting agents (colchicine, oryzaline or trifluralin) or spontaneously through unreduced gamete formation. However, the amphidiploid produced from *Erucastrum abyssinicum* × *B. oleracea* through colchicine treatment failed to produce any selfed or backcross progeny (Rao *et al.*, 1996). It may be attributable to the presence of some sterility genes (prevalent in rice crop), normally associated with reproductive isolation.

HYBRID NECROSIS Symptoms of stunted growth, wilting, chlorosis and lethality are common in wide hybrids because of epistatic interactions between resistance R genes of the crossed species. This may activate autoimmune-like responses, leading to hybrid weakness (Bomblies *et al.*, 2007; Jeuken *et al.*, 2009). Negative interaction between *Hwi1* in wild rice (*Oryza rufipogon*) and *Hwi2* in indica rice (*Oryza sativa*), which encode two LRR-RLK proteins and putative subtilisin-like protease, respectively, is detected in interspecific hybrids, causing hybrid weakness. Earlier the Dobzhansky–Muller model also explained that F_1 inviability was caused by deleterious interactions between lineage-specific alleles at two or more loci (Dobzhansky, 1937; Muller, 1942). Use of a large number of accessions of parental species in cross combinations may be useful to overcome hybrid necrosis.

REDUCED CHROMOSOME PAIRING Homoeologous pairing is critical for genetic exchanges between crop and wild genomes. Such exchanges can be facilitated by suppressing or bypassing genetic factors that prevent chromosome pairing. This has been demonstrated in wheat by inducing translocations through X-rays or incorporation of mutants of *Ph1* locus. Homoeologous pairing between allohexaploid wheat and *Aegilops speltoides* (wild relative of wheat) could be facilitated by the deletion of *Ph1* locus (Riley and Chapman, 1958; Sears, 1976). A gene named *PrBn* has also been recognized as a significant gene governing homoeologous pairing in *B. napus*, but it displayed incomplete penetrance or variable expressivity (Jenczewski *et al.*, 2003; Liu *et al.*, 2006).

LINKAGE DRAG Even if introgressed, the alien segment is mostly associated with undesirable genes resulting in reduction of yield and/or fitness. As an example, introgression of a restorer gene conferring male fertility to Ogura CMS from *R. sativus* into *B. napus* carried along an unwanted gene responsible for seed glucosinolate synthesis, and it took extensive efforts to disrupt this linkage (Delourme *et al.*, 1995; Primard *et al.*, 2005). However, if two lines possessing different overlapping alien introgressed segments carrying the same target gene are identified using densely placed molecular markers, then these lines can be intercrossed to yield recombinant progeny carrying the small target introgressed segment but not the deleterious one. Introgressed genetic variation is of no use if its effects are negatively associated with some other desirable traits. As an example,

although transfer of *Lr47* gene from *Triticum speltoides* govering resistance to leaf rust into wheat resulted in increased grain and flour protein concentration, it also caused overall 3.8% reduction in grain yield of wheat (Brevis *et al.*, 2008). In the absence of pairing between crop and wild genomes, linkage drag caused by physical linkages can only be disrupted by heavy doses of gamma irradiations.

Genetic conduits for alien gene transfer

Interspecific or intergeneric hybridization

Partially fertile F_1 hybrids can be directly backcrossed to cultivated species to facilitate transfer of alien chromosome segments into the cultivated genome. As an example, F_1 interspecific hybrids were synthesized through hybridization between *B. carinata* (BBCC, 2n = 34) and *B. oleracea botrytis* group (CC, 2n = 18) using ovule culture, with the purpose of incorporating resistance against black rot disease from *B. carinata* into *B. oleracea botrytis.* Hybrids were partially fertile and these could be backcrossed to recurrent species to introgress resistance against blackleg (Sharma *et al.*, 2017). Bang *et al.* (2007) backcrossed the intergeneric hybrids synthesized from reciprocal crosses of *B. oleracea* (CC, 2n = 18) and *Moricandia arvensis* (MaMa; 2n = 28) directly to *B. oleracea* to obtain sesquidiploid plants in BC_1 and BC_2 generations. In another study, semi-fertile F_1 plants containing 30 chromosomes consisting of two *B. oleracea* chromosome sets and one *Si. alba* chromosome set were recognized from intergeneric sexual hybridization between paternal *B. oleracea* var. *alboglabra* (CC, 2n = 18) and maternal *Si. alba* (SS, 2n = 24). Semi-fertile F_1 plants were backcrossed to *B. oleracea*, resulting in one monosomic alien addition line (Wei *et al.*, 2006).

Synthetic amphidiploids

An alternative to alien genetic introgressions is the production of fertile amphidiploids by inducing chromosome doubling. Consequent fertile amphidiploids can be used as bridging species and are backcrossed with recipient *Brassica* species to obtain descendants having additional/substituted alien chromosomes or segments. As 45 out of 63 cytodemes of *Brassica* coenospecies contain diploid wild species, it is ideal to first synthesize allotetraploids between wild species and diploid Brassicas. Many fertile amphidiploids between crop cultivars and wild species have been produced. These include: *B. fruticulosa* × *R. sativus* (Bang *et al.*, 1997, 2000), *B. maurorum* × *R. sativus* (Bang *et al.*, 1997, 1998), *Brassica oxyrrhina* × *R. sativus* (Bang *et al.*, 1997; Matsuzawa *et al.*, 1997), *B. rapa* × *Diplotaxis tenuifolia* (Jeong *et al.*, 2009), *B. rapa* × *Erucatrum cardaminoides*, *B. nigra* × *Erucastrum cardaminoides* (Chandra *et al.*, 2004a) and *Erucastrum canariense* × *B. rapa* (Bhaskar *et al.*, 2002). These amphidiploids exhibit low pollen and seed viability because of chromosomal rearrangements resulting from homoeologous pairing in earlier generations. However, in later generations, they can attain high pollen and seed fertility as a result of complete meiotic stabilization. The amphidiploids serve as permanent entities and can be maintained across generations if they remain meiotically stable. Such alloploids can be used directly as a source of alternate diversity or these can be used as a genetic conduit between a wild species and the recipient allotetraploid needing to be improved. Synthetic amphiploids benefit from fixed heterosis and provide other merits to create novel traits over parental species, such as increased biomass (Liu *et al.*, 2002; Qian *et al.*, 2003; Bansal *et al.*, 2012), seed yield (Osborn, 2004), sclerotinia resistance (Zhao *et al.*, 2006) and loss of self-incompatibility (Okamoto *et al.*, 2007). Some new crop species, such as 'Hakuran' (*B. campestris* × *B. oleracea*), 'Radicole' (*R. sativus* × *B. oleracea*) and 'Raparadish' (*B. campestris* × *R. sativus*) have also been established from synthetic amphiploid lines (Namai, 1987). Jesske *et al.* (2013) resynthesized 71 lines of *B. napus* involving wild *B. oleracea* or domesticated *B. oleracea* as one of the parents and demonstrated broad genetic diversity in resynthesized lines, which was absent in the breeding material of *B. napus.* Although resynthesized lines from wild *B. oleracea* were lower in yield than from domesticated *B. oleracea*, they produced high-yielding hybrids when crossed with adaptive genotypes of *B. napus.*

Monosomic and disomic alien addition lines

Recurrent backcrossing of F_1 hybrids or synthetic amphidiploids to the parental elite species

results in production of monosomic alien addition lines (MAALs) containing addition of an added chromosome from the alien species to the chromosome complement of cultivated species. MAALs are excellent genetic stocks for elucidating genome structure of wild species by dissecting these into individual chromosome units. These lines allocate alien genes to individual donor chromosome(s) and facilitate study of syntenic relationships between alien chromosome and respective orthologous recipient genome through intergenomic recombination (Namai, 1987; McGrath and Quiros, 1990; Prakash and Chopra, 1990; Matsuzawa *et al*., 1996). Selfing of monosomic alien addition lines precedes construction of disomic alien addition lines. An ideal set of monosomic alien addition lines represents the entire chromosome complement of a wild species individually in the background of recipient genome. To categorize them, unique morphological features of each alien chromosome or appropriate set of *in situ* probes or molecular markers specific for each pair of donor chromosomes are essential to assist in selection. *Brassica napus*, with extra chromosomes from *Diplotaxis muralis* (Fan *et al*. 1985), and *B. oleracea*, with added chromosomes from *R. sativus* (Kaneko *et al*. 1987), were first among many MAALs produced (Table 19.3). Delourme *et al*. (1989) were able to identify monosomic and disomic addition lines of *B. napus* with the extra chromosome from *Di. erucoides* in the BC_3 generation. These were later selfed to produce disomic addition plants. Akaba *et al*. (2009a,b) produced MAALs of *B. napus* with individual chromosomes of eight types (a to i, except for h type) of *R. sativus* (2n =18; RR), following sexual hybridization. Recently, Kang *et al*. (2014) reported a complete set of monosomic alien addition lines of *B. napus* with one of seven chromosomes of *I. indigotica* (Chinese woad; 2n = 14, II). Fortunately, one of the MAALs with cytoplasm of *I. indigotica* became available, which successfully restored fertility to the *inap* (from *I. indigotica*) male-sterile line of *B. napus* with carpelloid stamens and the line was subsequently selfed to produce rapeseed type plant (2n = 38) having one dominant restorer gene.

Chromosome substitution lines

Replacement of one chromosome pair of cultivated species with the homoeologous chromosome pair of wild species promotes synthesis of alien chromosome substitution lines. Generally, disomic alien chromosome substitutions could be obtained only for the corresponding homoeologous cultivated *Brassica* chromosome pair that can complement each other. There is no evidence of a whole set of chromosome substitution lines in Brassicaceae. Banga (1988) confirmed by analysing meiotic configurations that C genome chromosome substitutions spontaneously originated in progenies (F_7 generation) of an interspecific cross between *B. juncea* and *B. napus*. Gupta *et al*. (2016) provided molecular evidence for three whole chromosome substitutions and 13 major C genome segmental substitutions randomly replacing B genome chromosomes in derived *B. juncea* synthesized from non-parental digenomic species *B. napus* and *B. carinata*. Substitution plants with one to four pairs of chromosomes from *O. violaceus* were also identified from mixoploid hybrids of *B. carinata* and *B. juncea* with *O. violaceus* (Li *et al*., 2003).

Alien introgression lines

Introgression lines (ILs) carrying segmental chromosome substitutions or translocations, which integrate small alien chromosome segments containing the gene of interest, represent a more desirable approach to minimizing linkage drag, when compared with the addition or substitution of whole genome or a chromosome from a donor species. The substituted segments may result from meiotic recombination between cultivated genomes and their homoeologous counterparts from wild species as well as from spontaneous or induced translocations. Backcrossing of MAALs to the recipient parent, followed by consecutive selfing, results in plants containing short, overlapping introgressions, which cover a large proportion of the donor genome. ILs have become a valuable experimental material for molecular breeding and could be used to assess the action and interaction of genes across multiple years and in multiple site experiments. Although a countable number of alien addition lines has been developed in the genomic background of cultivated *Brassica* species and some of them have been successfully utilized to transfer alien genes to euploid karyotypes, e.g. incorporation of fertility restoration genes from *Mo. arvensis* into *B. juncea* (Prakash *et al*., 1998), from *E. lyratus*

Table 19.3. Monosomic and disomic alien addition lines developed in crop *Brassica.*

Recipient species	Donor species	MAALs obtained	Mode of synthesis	Reference
B. campestris syn. *B. rapa* (2n = 20)	*B. oleracea* (2n = 18)	8 (Monosomic)	Sexual hybridization	Quiros *et al*. (1987)
	B. alboglabra (2n = 18)	1 (Monosomic)	Sexual hybridization	Chen *et al*. (1992)
	B. alboglabra (2n = 18)	4 (Monosomic)	Sexual hybridization	Chen *et al*. (1997)
	B. oxyrrhina (2n = 18)	7 (Monosomic)	Sexual hybridization	Srinivasan *et al*. (1998)
	Moricandia arvensis (2n = 28)	1 (Monosomic)	Sexual hybridization	Tsutsui *et al*. (2011)
	B. oleracea (2n = 18)	9 (Monosomic)	Sexual hybridization	Heneen *et al*. (2012)
	B. oleracea var. *capitata* (2n = 18)	5 (Monosomic); 5 (disomic)	Sexual hybridization	Gu *et al*. (2015)
B. oleracea (2n = 18)	*Sinapsis alba* (2n = 24)	1 (Monosomic)	Sexual hybridization	Wei *et al*. (2006)
	Mo. arvensis (2n = 28)	1 (Monosomic)	Sexual hybridization	Bang *et al*. (2007)
	B. nigra (2n = 16)	7 (Monosomic)	Sexual hybridization	Tan *et al*. (2017)
	B. campestris (2n = 20)	4 (Monosomic)	Sexual hybridization	Li *et al*. (2013)
B. napus (2n = 38)	*Diplotaxis muralis* L. (2n = 42)	Monosomic addition lines (not characterized)	Sexual hybridization	Fan *et al*. (1985)
	B. nigra (2n = 16)	Disomic addition lines (not characterized)	Sexual hybridization	Jahier *et al*. (1989)
	Diplotaxis erucoides (2n = 14)	Monosomic and disomic addition lines (not characterized)	Sexual hybridization	Delourme *et al*. (1989)
	Sinapsis arvensis (2n = 18)	1 (Disomic)	Somatic hybridization	Hu *et al*. (2002a)
	Crambe abyssinicum (2n = 90)	2 (Monosomic); 2 (disomic)	Somatic hybridization	Wang *et al*. (2003, 2006)
	A. thaliana (2n = 10)	1 (Monosomic); 1 (disomic)	Somatic hybridization	Leino *et al*. (2004)
	Raphanus sativus (2n = 18)	9 (Disomic)	Sexual hybridization	Peterka *et al*. (2004); Budahn *et al*. (2008)
	Si. alba (2n = 24)	7 (Monosomic)	Somatic hybridization	Wang *et al*. (2005a,b)
	Orychophragmus violaceus (2n = 24)	8 (Monosomic)	Somatic hybridization	Zhao *et al*. (2008); Ding *et al*. (2013)
	R. sativus (2n = 18)	9 (Monosomic)	Sexual hybridization	Akaba *et al*. (2009a)
	B. juncea (2n = 36)	7 (Monosomic)	Sexual hybridization	Takashima *et al*. (2012)
	Isatis indigotica (2n = 14)	7 (Monosomic); 1 (disomic)	Somatic hybridization	Kang *et al*. (2014)

Continued

Table 19.3. Continued.

Recipient species	Donor species	MAALs obtained	Mode of synthesis	Reference
R. sativus (2n = 18)	*B. oleracea* (2n = 18)	7 (Monosomic)	Sexual hybridization	Kaneko *et al*. (1987)
	Mo. arvensis (2n = 28)	12 (Monosomic)	Sexual hybridization	Bang *et al*. (2002)
	B. rapa (2n = 20)	8 (Monosomic)	Sexual hybridization	Kaneko *et al*. (2001)

MAAL, monosomic alien addition line.

into *B. juncea* (Banga *et al*., 2003b) and from *I. indigotica* into *B. napus* (Li *et al*., 2019), a very limited set of recombinant lines have been reported, namely B genome introgression lines of *B. napus* (Dhaliwal *et al*., 2017), *B. napus–Si. alba* introgression lines (Li *et al*., 2012), *B. juncea–B. fruticulosa* introgression lines (Atri *et al*., 2012) and *B. juncea–Erucastrum cardaminoides* introgression lines (Rana *et al*., unpublished) that are well characterized using cytogenetic and molecular techniques.

Cytogenetic and Molecular Characterization of Wide Hybrids and Backcross Derivatives

Successful introgressive breeding depends upon the ability to trace introgression of alien genome from F_1 hybrids onwards in alien addition and recombinant lines. The purpose always is to develop improved cultivars with targeted introgression of alien genes for economically important traits with minimal linkage drag. It is mostly difficult to determine true hybridity of F_1 on the basis of morphological assessment of leaves, flowers and pods, as these are often more akin to either of the parents than to the expected intermediate phenotype. Conventional cytological tools can confirm the F_1 hybridity from reduced pollen fertility and occurrence of univalents, bivalents, multivalents, bridges and laggards in meiotic configurations. These, however, cannot uniequivocally discriminate between autosyndetic and allosyndetic pairing. It is also difficult to estimate the proportion of alien chromatin in backcross progenies. *In situ* hybridization techniques, however, allow microscopic visualization of complementary sequences on chromosome preparations using radioactive or fluorophores-labelled molecular probes, with excellent accuracy (Schwarzacher *et al*., 1992; Thomas *et al*., 1994; Chang and de Jong, 2005). Use of *in situ* hybridization also enables differentiation of autosyndetic and allosyndetic pairing in hybrids (Ge *et al*., 2009; Tu *et al*., 2009). Singh *et al*. (2016) demonstrated the frequent occurrence of two allopairs between A/C genome of *B. napus* and RR genome of *Raphanus raphanistrum* in F_1 hybrids using GISH analysis, indicating the potential for gene transfer from *R. raphanistrum* to *B. napus*. Fahleson *et al*. (1997) could detect one or two complete *Er. sativa* chromosomes in somatic hybrid progeny of *B. napus* and *Er. sativa* using differentially labelled DNA from two parental species. The complete chromosome complement of parental species was identified using whole-genome labelled probes in intertribal somatic hybrids of *R. sativus* and *B. rapa* with *I. indigotica* (Tu *et al*., 2008). Feng *et al*. (2009) determined the physical location of fertility restorer gene (*Rfo*) in canola by simultaneously using two bacterial artificial chromosome (BAC) clones flanking the *Rfo* gene, besides 45SrDNA sequences as molecular probes in a dual colour fluorescent *in situ* hybridization (FISH) experiment. Intergenomic association between *B. napus* and *Si. alba* chromosomes was effectively detected in backcross progenies of wide hybrids between *B. napus* ssp. *napus* L. (AACC, 2n = 38) and *Si. alba* L. (SS, 2n = 24) using whole-genomic DNA of *Si. alba* as a molecular probe (Wang *et al*., 2005b). Next-generation sequencing technologies are further helping in synthesizing robust and cost-effective, pooled oligo probes specific to each chromosome of a species whose genome has been sequenced using bioinformatic approaches. These can be used in multicolour FISH experiments to physically recognize the exact chromosomes

involved in allosyndetic or autosyndetic pairing and number, size and approximate location of alien chromosomes or segments in chromosome spreads of backcross progenies of wide hybrids (Guo *et al.*, 2015). However, such techniques cannot unravel small-sized or complex rearrangements. These also do not permit identification of chromosome breakpoints at nucleotide level in introgression lines. Diverse molecular-marker technologies, including restriction fragment-length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment-length polymorphism (AFLP), simple sequence repeats (SSRs) and SNP arrays, have been used to identify the proportion of alien genome in hybrids and backcross progenies and subsequent mapping of QTL from alien genome. Akaba *et al.* (2009a,b) clearly classified the MAALs by identifying all chromosomes of *R. sativus* obtained in BC_2 generation of intergeneric hybrid between synthetic amphiploids (RRAA and RRCC) and *B. napus* (AACC) using alien chromosome-specific RAPD markers. Atri *et al.* (2012) characterized the introgression lines of *B. juncea* harbouring alien segments from *B. fruticulosa* using SSR markers and identified the mean proportion of 49% and 35% of recipient and donor genome in substitution lines, respectively. SSR genotyping using B genome-specific markers, along with SNP genotyping (60k AC-SNP array), allowed the identification of B genome segments from B4, B6 and B7 introgressed into ten chromosomes of *B. napus* in 17 lines (Dhaliwal *et al.*, 2017). However, low placement of these markers or even paucity of alien chromosome-specific molecular markers limits the use of these marker-assisted technologies. These technologies fail to precisely delineate introgressed segments and identify underlying candidate genes. As the selection is based only on a few markers, usually desirable smaller introgressions may have escaped detection or broken linkage of marker with trait may create problems in marker-assisted selection for traits. Newer approaches, such as next-generation genome sequencing technologies, now permit identification of SNPs evenly dispersed all through the genome. These can help to saturate mapping populations by detecting all of the recombination events. Tanksley and Nelson (1996) proposed the advanced backcross QTL (AB-QTL) approach for mapping and transfer of desired QTL from wild germplasm into selected breeding materials. This approach also permits QTL analysis in advanced segregating backcrossed (BC_2F_2 or BC_2F_3) populations. In the advanced generations, most of the recipient parental genome is recovered/improved and any QTL/genes detected are free from the epistatic interactions implicated by the donor genome. More than 37 QTL for 11 fruit quality traits (fruit weight, dry matter weight, external colour, internal colour, locule number, wall thickness, firmness, fruit shape, stem scar, soluble solids content and pH) were identified (Celik *et al.*, 2017) in the inbred backcross population (BC_2F_6) developed through advanced backcross QTL strategy. This was facilitated by the use of interspecific SNPs between the wild species *S. pimpinellifolium* (LA1589) and *S. lycopersicum* cv. Tueza genomes. Aflitos *et al.* (2015) developed a novel bioinformatic tool (known as introgression browser) to visualize introgressions at nucleotide level and demonstrated the capability of the tool by identifying alien DNA in a panel of closely related *S. pimpinellifolium* by examining phylogenetic relationships of the introgressed segments in tomato. Cleavenger *et al.* (2017) developed an automated pipeline, IntroMap, for high resolution fine-mapping of alien introgressions by generating new diagnostic SNPs from any species/accession of interest and showed the efficiency of software in detecting alien introgressions in cultivated *Arachis hypogea* using SNP sets for five diploid wild *Arachis* species.

Case Studies for Successful Use of *Brassica* Wild Relatives for Germplasm Enhancement

Development of CMS (cytoplasmic male-sterile) and fertility restorer lines

Synthesis of novel CMS systems of alloplasmic origin by transferring the nucleus of cultivated species into cytoplasm of diverse wild species through successive backcrossing of sexual or somatic hybrids with the cultivated species (Warwick and Black, 1991; Pradhan *et al.*, 1992) is the best example of the use of alien genetic variation. Most alloplasmic lines are male-sterile because of nuclear-cytoplasmic incompatibility between wild mitochondrion and crop nuclear genome. CMS is the most efficient method of

pollination control for F_1 hybrid breeding (Tingdong *et al.*, 1990; Jain *et al.*, 1994) of all other systems, such as self-incompatibility and genetic male sterility. Ogura CMS (Ogura 1968) is the first alloplasmic CMS that is still being utilized worldover for F_1 hybrid breeding of *B. napus*, *B. juncea*, *B. oleracea* and *R. sativus.* It was first recognized in Japanese radish and subsequently incorporated into European radish and other cultivated Brassicas (Bannerot *et al.*, 1974; Heyn, 1976). Lines that depict stable expression of male sterility spanning different environments are used as maternal parents in crossing with restorer lines (that confer fertility to sterile lines) for commercial hybrid seed production. CMS sources from many wild species, such as *A. thaliana, O. violaceus, Mo. arvensis, Di. muralis, Si. arvensis, Si. alba, I. indigotica, Trachystoma ballii* and *E. lyratus* (Banuelos *et al.*, 2013; Banga *et al.*, 2015), have been developed in crop *Brassica* (Table 19.4). More recently, a novel CMS source, expressed as rudimentary anthers, was identified following sexual hybridization between *B. fruticulosa* and *B. juncea* (Atri *et al.*, 2016). Somatic hybridization with *I. indigotica* (Chinese woad), followed by recurrent backcrossing, also resulted in generation of a novel CMS source of *B. napus*. In this instance, tetradynamous stamens were transformed into carpelloid structures with stigmatoid tissues at their tips and ovule-like tissues in the margins, and the two shorter stamens into filaments without anthers (Du *et al.*, 2009; Kang *et al.*, 2017). As in alloplasmic lines produced through sexual hybridization, cytoplasm is contributed only by the wild female parent, many alloplasmic CMS lines show adverse effects because of incongruities between alien cytoplasm and crop nucleus. These include leaf chlorosis and poor female fertility. Somatic cell fusion is the only effective approach to rectifying these errors by facilitating recombinations between mitochondria of participating species, yielding several different alloplasmic lines with different mt-genome constitutions. Somatic hybridization also permits substitution of wild species' chloroplasts with those from recipient crop species. This helps to overcome leaf chlorosis evident in many alloplasmic lines. Low temperature chlorosis for ogura CMS was corrected through somatic hybridization (Pelletier *et al.*, 1983). This was possible by selecting progeny having choloroplast from *B. napus* and mitochondria from *R. sativus*, resulting from protoplast fusion between alloplasmic *B. napus* with normal *B. napus* (Pelletier *et al.*, 1983; Menczel *et al.*, 1987; Jarl and Bornman, 1988). Similarly, improvement of CMS line of *B. juncea* with Ogura cytoplasm, which had flowers with petaloid anthers and poor female fertility, has been reported. Protoplast fusion allowed mitochondrial recombination and a fully female-fertile CMS line was obtained (Kirti *et al.*, 1993, 1995b,c). To exploit CMS commercially for hybrid seed production, the exploration of fertility restorer genes capable of providing male fertility to CMS lines is essential. For alloplasmic CMS lines, *Rf* genes (restorer of fertility) are not generally available in crop germplasm. These need to be introgressed from cytoplasm donor wild species. *Rf* genes have been successfully introgressed from other species into cultivated *Brassica* species for the CMS-inducing cytoplasms, such as Ogura (Paulman and Robbelen, 1988; Sakai *et al.*, 1996), *T. ballii* (Kirti *et al.*, 1997), *Mo. arvensis* (Kirti *et al.*, 1998), *Erucastrum canariense* (Prakash *et al.*, 2001), *E. lyratus* (Banga *et al.*, 2003b; Deol *et al.*, 2003; Janeja *et al.*, 2003), *Brassica tournefortii* (Stiewe and Robbelen, 1994), *B. fruticulosa* (Atri *et al.*, 2016) and *I. indigotica* (Li *et al.*, 2019). Interestingly, cytoplasm of diverse wild species (*Diplotaxis berthautii, Di. catholica, Di. erucoides, Mo. arvensis, B. oxyrrhina, Si. alba, E. lyratus* and *D. tenuisilquae*) conferring CMS to cultivated *Brassica* possesses the same mitochondrial gene (*orf108*), suggesting a common molecular mechanism underlying CMS in these species, and a single *Rf* gene isolated from *Mo. arvensis* could restore the fertility of four CMS systems, namely, *Mo. arvensis, Di. catholica, Di. erucoides* and *Di. berthautii* (Bhat *et al.*, 2005, 2006, 2008). This implied coevolution of the same fertility-restoration system in these species (Ashutosh *et al.*, 2008; Kumar *et al.*, 2012). It is important to decipher the molecular mechanism governing each male sterility and fertility-restoration system to identify its uniqueness to reduce the chances of genetic vulnerability because of the use of a single cytoplasm.

Disease resistance

Brassica crops are suceptible to many diseases, such as sclerotinia stem rot, caused by *Sclerotinia*

Table 19.4. Cytoplasmic male-sterile sources in *Brassica.*

Cultivated species	Wild species	Mode of synthesis	Reference
B. rapa	Refined CMS (*ogura*) *B. napus*	Sexual hybridization	Delourme *et al*. (1994)
	Brassica oxyrrhina	Sexual hybridization	Prakash and Chopra (1988)
	Diplotaxis muralis	Sexual hybridization	Hinata (1979)
	Enarthocarpus lyratus	Sexual hybridization	Deol *et al*. (2003)
	Moricandia arvensis	Sexual hybridization	Tsutsui *et al*. (2011)
	Eruca sativa	Sexual hybridization	Matsuzawa *et al*. (1999)
B. oleracea	*Raphanus sativus*	Sexual hybridization	Bannerot *et al*. (1974)
		Protoplast fusion	Pelletier *et al*. (1989); Kao *et al*. (1992); Walters *et al*. (1992)
	Di. muralis	Sexual hybridization	Shinada *et al*. (2006)
B. juncea	*Brassica tournefortii*	Spontaneous	Rawat and Anand (1979); Pradhan *et al*. (1991)
	Refined CMS (*ogura*) *B. napus*	Sexual hybridization	Delourme *et al*. (1994)
	B. oxyrrhina	Sexual hybridization	Prakash and Chopra (1990)
	Diplotaxis erucoides	Sexual hybridization	Malik *et al*. (1999); Bhat *et al*. (2006)
	E. lyratus	Sexual hybridization	Banga *et al*. (2003b)
	Diplotaxis berthautii	Sexual hybridization	Malik *et al*. (1999); Bhat *et al*. (2008)
	Mo. arvensis	Protoplast fusion	Prakash *et al*. (1998)
	Diplotaxis siifolia	Sexual hybridization	Rao *et al*. (1994)
	Erucastrum canariense	Sexual hybridization	Prakash *et al*. (2001)
	Diplotaxis catholica	Sexual hybridization	Pathania *et al*. (2003)
		Protoplast fusion	Pathania *et al*. (2007)
	Trachystoma ballii	Protoplast fusion	Kirti *et al*. (1995b)
	R. sativus	Sexual hybridization	Bannerot *et al*. (1974); Paulmann and Robbelen (1988)
		Protoplast fusion	Pelletier *et al*. (1983); Sakai and Imamura (1990)
	B. tournefortii	Sexual hybridization	Mathias (1985)
		Protoplast fusion	Stiewe and Robbelen (1994); Liu *et al*. (1996)
	Di. muralis	Sexual hybridization	Pellan-Delourme and Renard (1987)
	Di. siifolia	Sexual hybridization	Rao and Shivanna (1996)
B. napus	*E. lyratus*	Sexual hybridization	Janeja *et al*. (2003)
	Sinapsis arvensis	Protoplast fusion	Hu *et al*. (2004)
	Arabidopsis thaliana	Protoplast fusion	Leino *et al*. (2003)
	Orychophragmus violaceus	Protoplast fusion	Mei *et al*. (2003)
	Sinapsis alba	Protoplast fusion	Wang *et al*. (2014)
	Isatis indigotica	Protoplast fusion	Du *et al*. (2009); Kang *et al*. (2014)
R. sativus	*Brassica maurorum*	Sexual hybridization	Bang *et al*. (2011)

sclerotiorum, blackleg (*Leptosphaeria maculans*), white rust (*Albugo candida*), alternaria blight (*Alternaria* spp.), downy mildew (*Peronospora parasitica*) and black rot (*Xanthomonas campestris pv. campestris*). These attack the plants during several phases and cause serious yield losses, depending upon severity of attack. Wild species of Brassicaceae family form a rich reservoir of genes for resistance to these diseases, as described earlier in this chapter. For example, eight species (*B. desnottesii, Ca. sativa, Coincya pseuderucastrum, Di. berthautii, Di. catholica, D. cretacea, Di. erucoides* and *Erucastrum gallicum*) were reported to be completely resistant to leaf spot disease (Sharma *et al.*, 2002). Wide hybridizations between Brassicas and wild relatives have enabled introgression of resistance factors against some of the diseases by exploiting homoeologous recombination between cultivated and wild genomes. However, only a few wild species have been used for truly

introgressive breeding. Blackleg, caused by *L. maculans* (Desm.) Ces. Et de Not. (imperfect stage *Phoma lingam*), is a serious disease of *B. napus* in most canola-growing areas of Europe, Australia, Canada and China. Yield losses may vary from 13 to 50%, depending on the virulence of the prevalent pathogenic strain. *Brassica* species (*B. nigra, B. juncea, B. carinata*), carrying the B genome have been reported to possess complete resistance to this pathogen (Sjodin and Glemelius, 1989a,b). Roy (1984) successfully transferred the resistance against disease from *B. juncea* to *B. napus*, following sexual hybridization. Introgression of a part of the B-genome showing resistance to blackleg disease into *B. napus* background from *B. juncea* has been demonstrated recently (Rashid *et al*., 2018). Dixelius (1999) investigated inheritance of resistance to *L. maculans* in near-isogenic lines of *B. napus* derived from repeated backcrossing for seven generations of asymmetric somatic hybrids between *B. napus* + *B. nigra* and *B. napus* + *B. juncea*, with *B. napus*. Chevre *et al*. (1996, 1997) reported the creation of a recombinant line of *B. napus* showing regular meiotic behaviour, carrying monogenic resistance against the disease from *B. nigra*. Resistance to *P. lingam* has been reported in a recombinant line of *B. napus* derived from an interspecific hybrid between *B. napus* and *B. juncea* (Saal *et al*., 2004). Alternaria black spot is the disease of *Brassica* crops caused by a complex of *Alternaria* species, primarily by *Alternaria brassicae* (Berk.) Sacc., that reduces photosynthetic efficiency, accelerates senescence and causes premature pod shatter and shrivelled seeds of plants (Shrestha *et al*., 2000). Transfer of resistance to Alternaria blight has been reported from *Si. alba* (Chevre *et al*., 1991). Bacterial soft rot, caused by *Erwinia carotovora* subsp. *carotovora*, is a serious disease in Chinese cabbage (*B. rapa* L., *pekinensis* Group). Interspecific hybridization between Chinese cabbage and cabbage (*B. oleracea L., capitata* Group) has permitted the development of a relatively resistant cultivar of Chinese cabbage, 'Hiratsuka No. 1' (Shimizu *et al*., 1962). However, a more detailed study involved the transfer of resistance to sclerotinia stem rot from wild species (*E. cardaminoides* and *B. fruticulosa*) into *B. juncea* (Garg *et al*., 2010). Introgressive breeding was facilitated by the development of interspecific/intergeneric hybridization between *B. rapa*/*B. nigra* and *E. cardaminoides* and *B. fruticulosa* (Chandra *et al*., 2004a,b). Synthetic amphiploids, obtained after somatic doubling of F_1s, were used as a bridging species to incorporate resistance into *B. juncea.* Garg *et al*. (2010) reported a high level of resistance against the fungus *Sc. sclerotiorum* in the resulting recombinant (introgression) lines. Genome-wide association analysis of ILs using the genotyping-by-sequencing approach allowed identification of eight significant QTL present on chromosomes A02, A03, A06, B02, B03, B04 and B07 linked with disease resistance. Annotation of associated genomic regions indicated the role of the disease resistance protein (TIR-NBS-LRR class) family, the subtilase family, the leucin-rich protein kinase and peroxidase super-family proteins in explaining resistance responses (Rana *et al*., unpublished). Introgression lines of *B. juncea* carrying genomic segments from the wild species *B. fruticulosa* have also been identified to have a significant level of resistance against the disease. Initially, Rana *et al*. (2017) documented ten significant marker–trait associations for resistance using SSR-based association mapping. Studies were further extended by the same group through the genotyping-by-sequencing (GBS) approach and 13 significant loci on chromosomes A01, A03, A04, A05, A08, A09 and B05, explaining 7.34–16.04% of phenotypic variation, linked with disease-resistant families, have been identified (Atri *et al*., unpublished).

Pest resistance

Brassicas are infested by a number of insect pests, with no known source of resistance in crop germplasms. These include aphids, diamondback moth, painted bug, flea beetles, mustard sawfly and hairy caterpillar. Aphids (*Lipaphis erysimi* Kaltenbach, *Myzus persiceae* and *Brevicoryne brassicae*) are of global occurrence, causing very serious harm to crops either directly by feeding or indirectly by spreading plant viral diseases (Dawson *et al*., 1990). Depending upon the severity of aphid infestation and crop stage, damage to cruciferous crops may range from 10 to 90%. *Brassica fruticulosa*, a wild species recognized to possess a high level of resistance against *Li. erysimi* Kaltenbach, has been utilized to produce *B. juncea* ILs, which vary in

resistance to mustard aphid. (Kumar *et al*., 2011). Atri *et al*. (2012) reported high fertility and molecular evidence regarding the presence of *B. fruticulosa* chromatin substitution in these ILs using SSR markers and identified some lines showing consistent resistant reaction to mustard aphid across 2 years. Further studies by the same research group helped in the identification of QTL for aphid resistance using GWAS analysis with GBS data. An important gene *CAT 2* known to combat insect herbivory was recognized consistently across seasons on chromosome B02 (Kaur *et al*., unpublished).

Resistance to abiotic stresses

Climatic changes, such as increasing temperature and unpredictable precipitation pattern, along with the deteriorating edaphic conditions, are among the major reasons of yield plateau (Lobell and Gourdji, 2012; Kang and Banga, 2013). Keeping in mind the rising global population, it is important to design crops that are less demanding and can perform well under suboptimal growth conditions. As wild species (inc. Brassicaceae) are adapted to a variety of environmental conditions and perform stably under stress conditions, they can be exploited to breed for abiotic stress resistance (Banga and Kang, 2013). *B. napus* is extremely sensitive to pod shattering at harvest stage because of either adverse windy weather or disturbance of plant canopy with combine machinery. The *Brassica* species carrying the B genome (*B. nigra*, *B. juncea* and specifically *B. carinata*) contain a high level of resistance to the pod shatter trait (Navabi *et al*., 2010; Dhaliwal *et al*., 2017). Recently, Raman *et al*. (2017) revealed five statistically significant QTL (LOD ≥ 3) that are linked with pod shatter resistance on chromosomes B1, B3, B8 and C5 of *B. carinata*. *Brassica oleracea* generally needs a lengthier cold period of 6 or more weeks (Friend, 1985) in comparison to the 4–8 weeks of low temperatures (5°C) required by *B. rapa* varieties (with vernalization requirement) (Kim *et al*., 2007) to flower. A late-bolting recombinant line of *B. rapa* with a winter cropping pattern was produced by replacing its *FLC* alleles with *FLC* alleles of *B. oleracea* (BoFLC2) through interspecific hybridization, followed by repeated backcrossing to *B. rapa* for its stable year-round production as a vegetable crop (Shea *et al*., 2018). Other examples include introgressing genomic segments for incorporating resistance to auxinic herbicide (to control broadleaf weeds) from *B. kaber* to *B. juncea* and *B. rapa*, by conventional crossing, coupled with *in vitro* embryo-rescue techniques (Mithila and Hall, 2013).

Genes for quality traits

The *Brassica* breeding programmes also aim at modifying the fatty acid profile of seed oils for specific purposes. Oils with a high amount of erucic acid and other fatty acids, such as lauric and nervonic acids, are essential for industrial use, whereas oils with higher levels of oleic and linoleic acids and low levels of erucic and linolenic acids are important for improved edible quality. Asymmetric somatic hybrids between *B. napus* and *Cr. abyssinica* contained significantly greater amounts of erucic acid than *B. napus*, which is a result of the transfer of alien chromatin responsible for high erucic acid content from *Cr. abyssinica* into the *B. napus* genome (Wang *et al*., 2003). Li *et al*. (2012) introgressed the genes for yellow seed colour from *S. alba* into *B. napus* following somatic hybridization and subsequent backcrossing with *B. napus*. Broccoli genotypes with enhanced levels of glucoraphanin from a wild species, *B. villosa*, have been developed that have anticarcinogenic properties (Sarikamis *et al*., 2006). Zhang *et al*. (2013) introgressed increased levels of oleic and reduced glucosinolate contents from *O. violaceus*. They also introgressed gene(s) for tightly compressed branches, rigid and wooden main stem and double low quality of oil from *Cap. bursa-pastoris*. Shen *et al*. (2018) also reported QTL co-localized on two chromosomes of A genome (A02 and A07) for plant height, branch number and branch initiation height using SNPs in a doubled-haploid mapping population derived from the cross between one *Cap. bursa-pastoris* introgression line and *B. napus*.

Summary

Family Brassicaceae (Cruciferae) includes nine genera (*Diplotaxis, Brassica, Eruca, Erucastrum,*

Hirschfeldia, Coincya, Sinapis, Sinapidendron and *Trachystoma*) from subtribe Brassicinae, along with two genera from subtribe Raphaninae (*Raphanus, Enarthocarpus*) and three genera from subtribe Moricandiinae (*Rytidocarpus, Moricandia, Pseuderucaria*). Together, these form *Brassica* coenospecies. Genetic studies, including whole-genome sequencing, have revealed that speciation and diversification in Brassicaceae occurred in the aftermath of a shared whole-genome triplication event, which was followed by rediploidization and tetraploidization. *Brassica* is the most important genus, as it includes six closely related crop species. These are cultivated as edible oilseed, vegetable, condiment, fodder or industrial crops. Three of these are diploids (*B. rapa* [AA], *B. nigra* [BB], *B. oleracea* [CC]), which hybridized naturally to produce three allotetraploids (*B. carinata* [BBCC], *B. juncea* [AABB] and *B. napus* [AACC]), as depicted in the triangle of U. Phylogenetic relationships between *Brassica* and remaining genera of the family, are very complex, with strong but sometimes leaky reproductive boundries. This, coupled with the knowledge about the existence of a large reservoir of genetic variation in wild and weedy species, has prompted attempts at their utilization as genetic resources for the improvement of crop *Brassica* species. Embryo-rescue and protoplast fusion techniques have helped to produce wide hybrids, bypassing sexual constraints. Many cytoplasmic male sterility, fertility-restoring systems and introgression lines varying in resistance to existing pests and diseases have been developed and characterized. Linkage drag is a key limitation to commercial exploitation of these introgression conduits.

References

Adkins, S.W., Wills, D., Boersma, M., Walker, S.R., Robinson, G., *et al*. (1997) Weeds resistant to chlorsulfuron and atrazine from the north-east grain region of Australia. *Weed Research* 37, 343–349.

Aflitos, S.A., Sanchez-Perez, G., de Ridder, D., Fransz, P., Schranz, M.E., *et al*. (2015) Introgression browser: High-throughput whole-genome SNP visualization. *The Plant Journal* 82, 174–182.

Agnihotri, A., Shivanni, K.R., Lakshmikumaran, M.S. and Jagannathan, V. (1991) Micropropagation and DNA analysis of wide hybrids of cultivated *Brassica*. *Proceedings of the GCIRC 8th International Rapeseed Congress*, Saskatoon, Saskatchewan, Canada, 9–11 July 1991, p. 151.

Ahuja, I., Bhaskar, P.B., Banga, S.K., Banga, S.S. and Rakow, G. (2003) Synthesis and cytogenetic characterization of intergeneric hybrids of *Diplotaxis siifolia* with *Brassica rapa* and *B. juncea*. *Plant Breeding* 122, 447–449.

Akaba, M., Kaneko, Y., Ito, Y., Nakata, Y., Bang, S.W., *et al*. (2009a) Production and characterization of *Brassica napus – Raphanus sativus* monosomic addition lines mediated by the synthetic amphidiploid '*Raphanobrassica*'. *Breeding Science* 59, 109–118.

Akaba, M., Kaneko, Y., Hatakeyama, K., Ishida, M., Bang, S.W., *et al*. (2009b) Identification and evaluation of clubroot resistance of radish chromosome using a *Brassica napus–Raphanus sativus* monosomic addition line. *Breeding Science* 59, 203–206.

Ali, A., McLaren, R.D. and Machado, V.S. (1986) Chloroplastic resistance to triazine herbicides in *Sinapis arvensis* L. (wild mustard). *Weed Research* 26, 39–44.

Aminidehaghi, M., Rezaeinodehi, A. and Khangholi, S. (2006) Allelopathic potential of *Alliaria petiolata* and *Lepidium perfoliatum*, two weeds of the Cruciferae family. *Journal of Plant Diseases and Protection* 20, 455–462.

Anderson, M.D., Peng, C. and Weissl, M.J. (1992) *Crambe*, *Crambe abyssinica* Hochst., as a flea beetle resistant crop (Coleoptera: Chrysomelidae). *Journal of Economic Entomology* 85, 594–600.

Angelini, L.G., Moscheni, E., Colonna, G., Belloni, P. and Bonari, E. (1997) Variation in agronomic characteristics and seed oil composition of new oilseed crops in central Italy. *Industrial Crops and Products* 6, 313–323.

Apel, P., Horstmann, C. and Pfeffer, M. (1997) The *Moricandia* syndrome in species of the *Brassicaceae* – evolutionary aspects. *Photosynthetica* 33, 205–215.

Ashraf, M. (1994) Organic substances responsible for salt tolerance in *Eruca sativa*. *Biologia Plantarum* 36, 255–259.

Ashraf, M. and Noor, R. (1993) Growth and pattern of ion uptake in *Eruca sativa* Mill. under salt stress. *Angewandte Botanik* 67, 17–21.

Ashutosh., Kumar, P., Dinesh Kumar, V., Sharma, P.C., Prakash, S., *et al*. (2008) A novel *orf108* co-transcribed with the *atpA* gene is associated with cytoplasmic male sterility in *Brassica juncea* carrying *Moricandia arvensis* cytoplasm. *Plant and Cell Physiology* 49, 284–289.

Atri, C., Kumar, B., Kumar, H., Kumar, S., Sharma, S., *et al*. (2012) Development and characterization of *Brassica juncea – fruticulosa* introgression lines exhibiting resistance to mustard aphid (*Lipaphis erysimi* Kalt). *BMC Genetics* 13, 104. DOI: 10.1186/1471-2156-13-104.

Atri, C., Kaur, B., Sharma, S., Gandhi, N., Verma, H., Goyal, A., *et al*. (2016) Substituting nuclear genome of *Brassica juncea* (L.) Czern & Coss. in cytoplasmic background of *Brassica fruticulosa* results in cytoplasmic male sterility. *Euphytica* 209, 31–40. DOI: 10.1007/s10681-015-1628-4.

Badger, B. (1988) In search of a yellow evergreen azalea (how to hybridize for a yellow evergreen azalea). *Journal of the American Rhododendron Society* 42, 74–79.

Bang, S.W., Kaneko, Y. and Matsuzawa, Y. (1997) Production of new intergeneric hybrids between *Raphanus sativus* and *Brassica* wild species. *Japanese Journal of Breeding* 47, 223–228.

Bang, S.W., Kaneko, Y. and Matsuzawa, Y. (1998) Cytogenetical stability and fertility of an intergeneric amphidiploid line synthesized from *Brassica maurorum* Durieu and *Raphanus sativus* L. *Bulletin of the College of Agriculture, Utsunomiya University* 17, 23–29.

Bang, S.W., Kaneko, Y. and Matsuzawa, Y. (2000) Cytogenetical stability and fertility in an intergeneric amphidiploid line synthesized from *Brassica fruticulosa* Cyr. ssp. *mauritanica* (Coss.) Maire. and *Raphanus sativus* L. *Bulletin of the College of Agriculture, Utsunomiya University* 17, 67–73.

Bang, S.W., Kaneko, Y., Matsuzawa, Y. and Bang, K.S. (2002) Breeding of *Moricandia arvensis* monosomic chromosome addition lines (2n = 19) of alloplasmic (*M. arvensis*) *Raphanus sativus*. *Breeding Science* 52, 193–199.

Bang, S.W., Mizuno, Y., Kaneko, Y., Matsuzawa, Y. and Bang, K.S. (2003) Production of intergeneric hybrids between the C3-C4 intermediate species *Diplotaxis tenuifolia* (L.) DC. and *Raphanus sativus* L. *Breeding Science* 53, 231–236.

Bang, S.W., Sugihara, K., Jeung, B.H., Kaneko, R., Satake, E., *et al*. (2007) Production and characterization of intergeneric hybrids between *Brassica oleracea* and a wild relative *Moricandia arvensis*. *Plant Breeding* 126, 101–103.

Bang, S.W., Tsutsui, K., Shim, S. and Kaneko, Y. (2011) Production and characterization of the novel CMS line of radish (*Raphanus sativus*) carrying *Brassica maurorum* cytoplasm. *Plant Breeding* 130, 410–412.

Banga, S.K., Kumar, P., Bhajan, R., Singh, D. and Banga, S.S. (2015) Genetics and breeding. In: Kumar, A., Banga, S.S., Meena, P.D. and Kumar P.R. (eds) *Brassica Oilseeds: Breeding and Management*. CAB International, Wallingford, UK, pp. 11–41.

Banga, S.S. (1988) C-genome chromosome substitution lines in *Brassica juncea* (L.) Coss. *Genetica* 77, 81–84.

Banga, S.S. and Kang, M.S. (2013) Developing climate resilient crops: A conceptual framework. In: Kang, M.S. and Banga, S.S. (eds) *Combating Climate Change: An Agricultural Perspective*. CRC Press, Boca Raton, Florida, pp. 141–162.

Banga, S.S., Bhaskar, P.B. and Ahuja, I. (2003a) Synthesis of intergeneric hybrids and establishment of genomic affinity between *Diplotaxis catholica* and crop *Brassica* species. *Theoretical and Applied Genetics* 106, 1244–1247.

Banga, S.S., Deol, J.S. and Banga, S.K. (2003b) Alloplasmic male-sterile *Brassica juncea* with *Enarthrocarpus lyratus* cytoplasm and the introgression of gene(s) for fertility restoration from cytoplasm donor species. *Theoretical and Applied Genetics* 106, 1390–1395.

Bannerot, H., Boulidard, L., Cauderon, Y. and Tempe, J. (1974) Transfer of cytoplasmic male sterility from *Raphanus sativus* to *Brassica oleracea*. *Proceedings of Eucarpia Meeting Cruciferae* 25, 52–54.

Bansal, P., Banga, S. and Banga, S.S. (2012) Heterosis as investigated in terms of polyploidy and genetic diversity using designed *Brassica juncea* amphiploid and its progenitor diploid species. *PLoS ONE* 7, e29607. DOI: 10.1371/journal.pone.0029607.

Bansal, V.K., Tewari, J.P., Tewari, I., Gomez-Campo, C. and Stringam, G.R. (1997) Genus *Eruca*: A potential source of white rust resistance in cultivated brassicas. *Plant Genetic Resources Newsletter* 109, 25–26.

Banuelos, G.S., Dhillon, K.S. and Banga, S.S. (2013) Oilseed Brassicas. In: Singh, B.P. (ed.) *Biofuel Crops: Production, Physiology and Genetics*. CAB International, Wallingford, UK, pp. 339–368.

Batra, V., Prakash, S. and Shivanna, K.R. (1990) Intergeneric hybridization between *Diplotaxis siifolia*, a wild species and crop brassicas. *Theoretical and Applied Genetics* 80, 537–541.

Bauer-Weston, B., Keller, W., Webb, J. and Gleddie, S. (1993) Production and characterization of asymmetric somatic hybrids between *Arabidopsis thaliana* and *Brassica napus*. *Theoretical and Applied Genetics* 86, 150–158.

Bauwe, H. (1983) Comparative phylogenetic age of C3-C4 intermediate species of *Moricandia* determined by isoelectric focusing and amino acid composition of small subunit of ribulose 1,5-bisphosphate carboxylase oxygenase. *Photosynthetica* 17, 442–449.

Beckie, H.J., Hall, L.M., Tardif, F.J. and Seguin-Swartz, G. (2007) Acetolactate synthase inhibitor-resistant stinkweed (*Thlaspi arvense* L.) in Alberta. *Canadian Journal of Plant Science* 87, 965–972.

Bedinger, P.A., Broz, A.K., Tovar-Mendez, A. and McClure, B. (2017) Pollen-pistil interactions and their role in mate selection. *Plant Physiology* 173, 79–90.

Begum, F., Paul, S., Bag, N., Sikdar, S.R. and Sen, S.K. (1995) Somatic hybrids between *Brassica juncea* (L). Czern. and *Diplotaxis harra* (Forsk.) Boiss and the generation of backcross progenies. *Theoretical and Applied Genetics* 91, 1167–1172.

Bennett, M.D., Finch, R.A. and Barclay, I.R. (1976) The time rate and mechanism of chromosome elimination in *Hordeum* hybrids. *Chromosoma* 54, 175–200.

Bhaskar, P., Ahuja, I., Janeja, H. and Banga, S.S. (2002) Intergeneric hybridization between *Erucastrum canariense* and *Brassica rapa*. Genetic relatedness between EC and A genomes. *Theoretical and Applied Genetics* 105, 754–758.

Bhat, S.R., Prakash, S., Kirti, P.B., Kumar, V.D. and Chopra, V.L. (2005) A unique introgression from *Moricandia arvensis* confers male fertility upon two different cytoplasmic male-sterile lines of *Brassica juncea*. *Plant Breeding* 124, 117–120.

Bhat, S.R., Vijayan, P., Ashutosh., Dwivedi, K.K. and Prakash, S. (2006) *Diplotaxis erucoides* – induced cytoplasmic male sterility in *Brassica juncea* is rescued by the *Moricandia arvensis* restorer: Genetic and molecular analyses. *Plant Breeding* 125, 150–155.

Bhat, S.R., Kumar, P. and Prakash, S. (2008) An improved cytoplasmic male sterile (*Diplotaxis berthautii*) *Brassica juncea*: Identification of restorer and molecular characterization. *Euphytica* 159, 145–152.

Bing, D.J. (1991) Potential of gene transfer among oilseed *Brassica* and their weedy relatives. *Proceedings of the GCIRC 8th International Rapeseed Congress*, Saskatoon, Saskatchewan, Canada, 9–11 July 1991, pp. 1022–1027.

Boaz, M., Plitmann, U. and Heyn, C.C. (1990) The ecogeographic distribution of breeding systems in the Cruciferae (*Brassicaceae*) of Israel. *Israel Journal of Plant Sciences* 39, 31–42.

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., *et al*. (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biology* 5, e236, DOI: 10.1371/journal.pbio.0050236.

Boutsalis, P., Karotam, J. and Powles, S.B. (1999) Molecular basis of resistance to acetolactate synthase-inhibiting herbicides in *Sisymbrium orientale* and *Brassica tournefortii*. *Pesticide Science* 55, 507–516.

Boyd, R.S. and Martens, S.N. (1998) Nickel hyperaccumulation by *Thlaspi montanum* var. *montanum* (Brassicaceae): A constitutive trait. *American Journal of Botany* 85, 259–265.

Brevis, J.C., Chicaiza, O., Khan, I.A., Jackson, L., Morris, C.F., *et al*. (2008) Agronomic and quality evaluation of common wheat near-isogenic lines carrying the leaf rust resistance gene *Lr47*. *Crop Science* 48, 1441–1451.

Brewer, E.P., Saunders, J.A., Angle, J.S., Chaney, R.L. and McIntosh, M.S. (1999) Somatic hybridization between the zinc accumulator *Thlaspi caerulescens* and *Brassica napus*. *Theoretical and Applied Genetics* 99, 761–77.

Brun, H. and Tribodet, M. (1995) Pathogenicity of *Leptosphaeria maculans* isolates on one ecotype of *Arabidopsis thaliana*. *Cruciferae Newsletter* 17, 74–75.

Buckler, E.S., Thornsberry, J.M. and Kresovich, S. (2001) Molecular diversity, structure and domestication of grasses. *Genetics Research* 77, 213–218.

Budahn, H., Schrader, O. and Peterka, H. (2008) Development of a complete set of disomic rape-radish chromosome-addition lines. *Euphytica* 162, 117–128.

Burkart-Waco, D., Ngo, K., Dilkes, B., Josefsson, C. and Comai, L. (2013) Early disruption of maternal–zygotic interaction and activation of defense-like responses in *Arabidopsis* interspecific crosses. *The Plant Cell* 25, 2037–2055.

Bushell, C., Spielman, M. and Scott, R.J. (2003) The basis of natural and artificial postzygotic hybridization barriers in *Arabidopsis* species. *The Plant Cell* 15, 1430–1442.

Carcamo, H., Olfert, O., Dosdall, L., Herle, C., Beres, B., *et al*. (2007) Resistance to cabbage seedpod weevil among selected Brassicaceae germplasm. *The Canadian Entomologist* 139, 658–669.

Carlson, K.D. and Tookey, H.L. (1983) *Crambe* meal as a protein source for feeds. *Journal of the American Oil Chemists Society* 60, 1979–1985.

Celik, I., Gurbuz, N., Uncu, A.T., Frary, A. and Doganlar, S. (2017) Genome-wide SNP discovery and QTL mapping for fruit quality traits in inbred backcross lines (IBLs) of *Solanum pimpinellifolium* using genotyping by sequencing. *BMC Genomics* 18, 1. DOI: 10.1186/s12864-016-3406-7.

Chander, H. and Bakhetia, D.R.C. (1998) Evaluation of some cruciferous genotypes at seedling stage for resistance to mustard aphid, *Lipaphis erysimi* (Kalt.) under screen house and field conditions. *Journal of Insect Science* 11, 19–25.

Chandra, A., Gupta, M.L., Ahuja, I., Kaur, G. and Banga, S.S. (2004a) Intergeneric hybridization between *Erucastrum cardaminoides* and two diploid crop *Brassica* species. *Theoretical and Applied Genetics* 108, 1620–1626.

Chandra, A., Gupta, M.L., Banga, S.S. and Banga, S.K. (2004b) Production of an interspecific hybrid between *Brassica fruticulosa* and *B. rapa*. *Plant Breeding* 123, 497–498.

Chang, S.B. and de Jong, H. (2005) Production of alien chromosome additions and their utility in plant genetics. *Cytogenetic and Genome Research* 109, 335–343.

Chatterjee, G., Sikdar, S.R., Das, S. and Sen, S.K. (1988) Intergeneric somatic hybrid production through protoplast fusion between *Brassica juncea* and *Diplotaxis muralis*. *Theoretical and Applied Genetics* 76, 915–922.

Chaudhury, A.M., Ming, L., Miller, C., Craig, S., Dennis, E.S., *et al*. (1997) Fertilization-independent seed development in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the USA* 94, 4223–4228.

Chen, B.Y., Simonsen, V., Lanner-Herrera, C. and Heneen, W.K. (1992) A *Brassica campestris–alboglabra* addition line and its use for gene mapping, intergenomic gene transfer and generation of trisomics. *Theoretical and Applied Genetics* 84, 592–599.

Chen, B.Y., Cheng, B.F., Jørgensen, R.B. and Heneen, W.K. (1997) Production and cytogenetics of *Brassica campestris – alboglabra* chromosome addition lines. *Theoretical and Applied Genetics* 94, 633–640.

Chen, C.Y. and Seguin-Swarts, G. (1997) The use of a β-glucuronidase-marked isolate of *Leptosphaeria maculans* to study the reaction of crucifers to the blackleg fungus. *Canadian Journal of Plant Pathology* 19, 327–330.

Chen, C.Y. and Seguin-Swartz, G. (1999) Reaction of wild crucifers to *Leptosphaeria maculans*, the causal agent of blackleg of crucifers. *Canadian Journal of Plant Pathology* 21, 361–367.

Chen, H.F., Wang, H. and Li, Z.Y. (2007a) Intertribal crosses between *Brassica* species and *Capsella bursa-pastoris* for the improvement of oil quality and resistance to *Sclerotinia sclerotiorum* of *Brassica* crops. *Proceedings of the GCIRC 12th International Rapeseed Congress*, Wuhan, China, 26–30 March 2007, pp. 411–413.

Chen, H.F., Wang, H. and Li, Z.Y. (2007b) Production and genetic analysis of partial hybrids in intertribal crosses between *Brassica* species (*B. rapa, B. napus*) and *Capsella bursa-pastoris*. *Plant Cell Reports* 26, 1791–1800.

Chevre, A.M., Eber, F., Brun, H., Plessis, J., Primard, C., *et al*. (1991) Cytogenetic studies of *Brassica napus–Sinapsis alba* hybrids from ovary culture and protoplast fusion. Attempts to introduce *Alternaria* resistance into rapeseed. *Proceedings of the* GCIRC *8th International Rapeseed Congress*, Saskatoon, Saskatchewan, Canada, 9–11 July 1991, pp. 346–351.

Chevre, A.M., Eber, F., This, P., Barret, P., Tanguy, X., *et al*. (1996) Characterization of *Brassica nigra* chromosomes and of blackleg resistance in *B. napus–B. nigra* addition lines. *Plant Breeding* 115, 113–118.

Chevre, A.M., Barret, P., Eber, F., Dupuy, P., Brun, H., *et al*. (1997) Selection of stable *Brassica napus–B. juncea* recombinant lines resistant to blackleg (*Leptosphaeria maculans*). 1. Identification of molecular markers, chromosomal and genomic origin of the introgression. *Theoretical and Applied Genetics* 95, 1104–1111.

Christey, M.C. (2004) *Brassica* protoplast culture and somatic hybridization. Pua, E.C. and Douglas, C.J. (eds) *Biotechnology in Agriculture and Forestry*. Springer, New York, pp. 119–148.

Christey, M.C., Makaroff, C.A. and Earle, E.D. (1991) Atrazine-resistant cytoplasmic male-sterile-nigra broccoli obtained by protoplast fusion between cytoplasmic male-sterile *Brassica oleracea* and atrazine-resistant *Brassica campestris*. *Theoretical and Applied Genetics* 83, 201–208.

Chrungu, B., Verma, N., Mohanty, A., Pradhan, A. and Shivanna, K.R. (1999) Production and characterization of interspecific hybrids between *Brassica maurorum* and crop brassicas. *Theoretical and Applied Genetics* 98, 608–613.

Clevenger, J., Bertioli, D.J., Leal-Bertioli, S.C.M., Chu, Y., Stalker, H.T., *et al*. (2017) IntroMap: A pipeline and set of diagnostic diploid *Arachis* SNPs as a tool for mapping alien introgressions in *Arachis hypogaea*. *Peanut Science* 44, 66–73.

Cole, R.A. (1994) Isolation of a chitin-binding lectin, with insecticidal activity in chemically-defined synthetic diets, from two wild *Brassica* species with resistance to cabbage aphid *Brevicoryne brassicae*. *Entomologia Experimentalis et Applicata* 72, 181–187.

Conn, K.L. and Tewari, J.P. (1986) Hypersensitive reaction induced by *Alternaria brassicae* in *Eruca sativa*, an oil yielding crucifer. *Canadian Journal of Plant Pathology* 8, 348.

Conn, K.L., Tewari, J.P. and Dahiya, J.S. (1988) Resistance to *Alternaria brassicae* and phytoalexin-elicitation in rapeseed and other crucifers. *Plant Science* 56, 21–25.

Craig, A. and Millam, S. (1995) Modification of oilseed rape to produce oils for industrial use by means of applied tissue culture methodology. *Euphytica* 85, 323–327.

Dallavalle, E., Lazzeri, L. and Curto, G. (2005) Life cycle duration of *Meloidogyne incognita* and host status of *Brassicaceae* and *Capparaceae* selected for glucosinate content. *Nematology* 7, 203–212.

Daun, J., Barthet, V. and Scarth, R. (2003) Erucic acid levels in *Sinapis arvensis* L from different parts of the world. *Proceedings of the GCIRC 11th International Rapeseed Congress*, Copenhagen, Denmark, 6–10 July 2003, pp. 290–292.

Davies, D.R. (1974) Chromosome elimination in inter-specific hybrids. *Heredity* 32, 267–270.

Dawson, G.W., Griffiths, D.C., Merritt, L.A., Mudd, A., Pickett, J.A., *et al*. (1990) Aphid semiochemicals – a review, and recent advances on the sex pheromone. *Journal of Chemical Ecology* 16, 3019–3030.

De Candolle, A.P. (1821) Cruciferae. *Systema Natural* 2, 139–700.

Delourme, R., Eber, F. and Chevre, A.M. (1989) Intergeneric hybridization of *Diplotaxis erucoides* with *Brassica napus*. I. Cytogenetic analysis of F_1 and BC_1 progeny. *Euphytica* 41, 123–128.

Delourme, R., Eber, F. and Renard, M. (1994) Transfer of radish cytoplasmic male sterility from *Brassica napus* to *B. juncea* and *B. rapa*. *Cruciferae Newsletter* 16, 79.

Delourme, R., Eber, F. and Renard, M. (1995) Breeding double low restorer lines in radish cytoplasmic male sterility of rapeseed (*Brassica napus* L.). *Proceedings of the GCIRC 9th International Rapeseed Congress*. Cambridge, UK, pp. 6–8.

Deol, J.S., Shivanna, K.R., Prakash, S., Banga, S.S. and Robbelen, G. (2003) *Enarthrocarpus lyratus* – based cytoplasmic male sterility and fertility restorer system in *Brassica rapa*. *Plant Breeding* 122, 438–440.

Dhaliwal, I., Mason, A.S., Banga, S., Bharti, S., Kaur, B., *et al*. (2017) Cytogenetic and molecular characterization of B-genome introgression lines of *Brassica napus* L. *Genes, Genomes, Genetics* 7, 77–86.

Dierig, D.A., Shannon, M.C. and Grieve, C.M. (2001) Registration of WCL-SL1 salt tolerant *Lesquerella fendleri* germplasm. *Crop Science* 41, 604–604.

Dierig, D.A., Tomasi, P.M., Salywon, A.M. and Ray, D. T. (2004) Improvement in hydroxy fatty acid seed oil content and other traits from interspecific hybrids of three *Lesquerella* species: *Lesquerella fendleri*, *L. pallida*, and *L. lindheimeri*. *Euphytica* 139, 199–206.

Ding, L., Zhao, Z.G., Ge, X.H. and Li, Z.Y. (2013) Intergeneric addition and substitution of *Brassica napus* with different chromosomes from *Orychophragmus violaceus*: Phenotype and cytology. *Scientia Horticulturae* 164, 303–309.

Dixelius, C. (1999) Inheritance of the resistance to *Leptosphaeria maculans* of *Brassica nigra* and *B. juncea* in near-isogenic lines of *B. napus*. *Plant Breeding* 118, 151–156.

Dobzhansky, T. (1937) *Genetics and the Origin of Species*. Columbia University Press, New York.

Dresselhaus, T., Lausser, A. and Marton, M.L. (2011) Using maize as a model to study pollen tube growth and guidance, cross-incompatibility and sperm delivery in grasses. *Annals of Botany* 108, 727–737.

Du, X.Z., Ge, X.H., Yao, X.C., Zhao, Z.G. and Li, Z.Y. (2009) Production and cytogenetic characterization of intertribal somatic hybrids between *Brassica napus* and *Isatis indigotica* and backcross progenies. *Plant Cell Reports* 28, 1105–1113.

Duke, James A. (1983) *Handbook of Energy Crops*. Purdue University, Center for New Crops and Plant Products, West Lafayette, Indiana.

Dumas, C. and Knox, R.B. (1983) Callose and determination of pistil viability and incompatibility. *Theoretical and Applied Genetics* 67, 1–10.

Dushenkov, S., Skarzhinskaya, M., Glimelius, K., Gleba, D. and Raskin, I. (2002) Bioengineering of a phytoremediation plant by means of somatic hybridization. *International Journal of Phytoremediation* 4, 117–126.

Ellis, P.R. and Farrell, J.A. (1995) Resistance to cabbage aphid (*Brevicoryne brassicae*) in six *Brassica* accessions in New Zealand. *New Zealand Journal of Crop and Horticultural Science* 23, 25–29.

Ellis, P.R., Pink, D.A.C., Barber, N.E. and Mead, A. (1999) Identification of high levels of resistance to cabbage root fly, *Delia radicum*, in wild *Brassica* species. *Euphytica* 110, 207–214.

Escobar-Restrepo, J.M., Huck, N., Kessler, S., Gagliardini, V., Gheyselinck, J., *et al*. (2007) The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. *Science* 317, 656–660.

Fagbenro, O.A. (2004) Soybean meal replacement by roquette (*Eruca sativa* Miller) seed meal as protein feedstuff in diets for African Catfish, *Clarias gariepinus* (Burchell 1822), fingerlings. *Aquaculture Research* 35, 917–923.

Fahleson, J., Rahlen, L. and Glimelius, K. (1988) Analysis of plants regenerated from protoplast fusions between *Brassica napus* and *Eruca sativa*. *Theoretical and Applied Genetics* 76, 507–512.

Fahleson, J., Eriksson, I., Landgren, M., Stymne, S. and Glimelius, K. (1994a) Intertribal somatic hybrids between *Brassica napus* and *Thlaspi perfoliatum* with high content of the *T. perfoliatum*-specific nervonic acid. *Theoretical and Applied Genetics* 87(7), 795–804.

Fahleson, J., Eriksson, I. and Glimelius, K. (1994b) Intertribal somatic hybrids between *Brassica napus* and *Barbarea vulgaris* – production of *in vitro* plantlets. *Plant Cell Reports* 13(7), 411–416.

Fahleson, J., Lagercrantz, U., Mouras, A. and Glimelius, K. (1997) Characterization of somatic hybrids between *Brassica napus* and *Eruca sativa* using species-specific repetitive sequences and genomic *in situ* hybridization. *Plant Science* 123, 133–142.

Fan, Z., Tai, W. and Stefansson, B.R. (1985) Male sterility in *Brassica napus* L. associated with an extra chromosome. *Canadian Journal of Genetics and Cytology* 27, 467–471.

Faulkner, K., Mithen, R. and Williamson, G. (1998) Selective increase of the potential anticarcinogen 4-methylsulphinylbutyl glucosinolate in broccoli. *Carcinogenesis* 19, 605–609.

Feng, J., Primomo, V., Li, Z., Zhang, Y., Jan, C.C., *et al*. (2009) Physical localization and genetic mapping of the fertility restoration gene *Rfo* in canola (*Brassica napus* L.). *Genome* 52, 401–407.

Finch, R.A. (1983) Tissue-specific elimination of alternative whole parental genomes in one barley hybrid. *Chromosoma* 88, 386–393.

Forsberg, J., Landgren, M. and Glimelius, K. (1994) Fertile somatic hybrids between *Brassica napus* and *Arabidopsis thaliana*. *Plant Science* 95(2), 213–223.

Forsberg, J., Dixelius, C., Lagercrantz, U. and Glimelius, K. (1998a) UV dose-dependent DNA elimination in asymmetric somatic hybrids between *Brassica napus* and *Arabidopsis thaliana*. *Plant Science* 131, 65–76.

Forsberg, J., Lagercrantz, U. and Glimelius, K. (1998b) Comparison of UV light, X-ray and restriction enzyme treatment as tools in production of asymmetric somatic hybrids between *Brassica napus* and *Arabidopsis thaliana*. *Theoretical and Applied Genetics* 96, 1178–1185.

Friend, D.J.C. (1985) *Brassica*. In: Halevy, A.H. (ed.) *Handbook of Flowering*. CRC Press, Boca Raton, Florida, pp. 48–77.

Friesen, L.S. and Powles, S.B. (2007) Physiological and molecular characterization of atrazine resistance in a wild radish (*Raphanus raphanistrum*) population. *Weed Technology* 21, 910–914.

Fujimoto, R., Sugimura, T., Fukai, E. and Nishio, T. (2006) Suppression of gene expression of a recessive SP11/SCR allele by an untranscribed SP11/SCR allele in *Brassica* self-incompatibility. *Plant Molecular Biology* 61, 577–587.

Gaikwad, K., Kirti, P.B., Sharma, A., Prakash, S. and Chopra, V.L. (1996) Cytogenetical and molecular investigations on somatic hybrids of *Sinapis alba* and *Brassica juncea* and their backcross progeny. *Plant Breeding* 115, 480–483.

Garg, H., Banga, S., Bansal, P., Atri, C. and Banga, S.S. (2007) Hybridizing *Brassica rapa* with wild crucifers *Diplotaxis erucoides* and *Brassica maurorum*. *Euphytica* 156, 417–424.

Garg, H., Atri, C., Sandhu, P.S., Kaur, B., Renton, M., *et al*. (2010) High level of resistance to *Sclerotinia sclerotiorum* in introgression lines derived from hybridization between wild crucifers and the crop Brassica species *Brassica napus* and *B. juncea*. *Field Crops Research* 117, 51–58.

Gavloski, J.E., Ekuere, U., Keddie, A., Dosdall, L., Kott, L., *et al*. (2000) Identification and evaluation of flea beetle (*Phyllotreta cruciferae*) resistance within Brassicaceae. *Canadian Journal of Plant Science* 80, 881–887.

Ge, X.H., Wang, J. and Li, Z.Y. (2009) Different genome-specific chromosome stabilities in synthetic *Brassica* allohexaploids revealed by wide crosses with *Orychophragmus*. *Annals of Botany* 104, 19–31.

Gehringer, A., Friedt, W., Luhs, W. and Snowdon, R.J. (2006) Genetic mapping of agronomic traits in false flax (*Camelina sativa* subsp. *sativa*). *Genome* 49, 1555–1563.

Gisbert, C., Almela, C., Velez, D., Lopez-Moya, J.R., de Haro, A., *et al*. (2008) Identification of As accumulation plant species growing on highly contaminated soils. *International Journal of Phytoremediation* 10, 185–196. DOI: 10.1080/15226510801997457.

Gleba, Y.Y. and Hoffmann, F. (1980) 'Arabidobrassica': A novel plant obtained by protoplast fusion. *Planta* 149(2), 112–117.

Glimelius, K. (1999) Somatic hybridization. In: Gomez-Campo, C. (ed.) *Biology of Brassica Coenospecies*. Elsevier Science, Amsterdam, The Netherlands, pp. 107–148.

Goffman, F.D., Thies, W. and Velasco, L. (1999) Chemotaxonomic value of tocopherols in *Brassicaceae*. *Phytochemistry* 50, 793–798.

Gokavi, S.S., Malleshi, N.G. and Guo, M. (2004) Chemical composition of garden cress (*Lepidium sativum*) seeds and its fractions and use of bran as a functional ingredient. *Plant Foods for Human Nutrition* 59, 105–111.

Gomez-Campo, C. (1999a) *Biology of Brassica Coenospecies* (Vol. 4). Elsevier Science, Amsterdam, The Netherlands.

Gomez-Campo, C. (1999b) Seedless and seeded beaks in the tribe *Brassiceae*. *Cruciferae Newsletter* 21, 11–13.

Grosser, J.W. and Gmitter, F.G. (2011) Protoplast fusion for production of tetraploids and triploids: Applications for scion and rootstock breeding in citrus. *Plant Cell, Tissue and Organ Culture* 104, 343–357. DOI: 10.1007/s11240-010-9823-4.

Grossniklaus, U., Vielle-Calzada, J.P., Hoeppner, M.A. and Gagliano, W.B. (1998) Maternal control of embryogenesis by MEDEA, a polycomb group gene in *Arabidopsis*. *Science* 280, 446–450.

Gu, A.X., Shen, S.X., Wang, Y.H., Zhao, J.J., Xuan, S.X., *et al*. (2015) Generation and characterization of *Brassica rapa* ssp. *Pekinensis–B. oleracea* var. *capitata* monosomic and disomic alien addition lines. *Journal of Genetics* 94, 435–444.

Guan, C., Li, F., Li, X., Chen, S., Liu, Z., *et al*. (2004) Resistance of rocketsalad (*Eruca sativa* Mill.) to stem rot (*Sclerotinia sclerotiorum*). *Zhongguo nongye kexue* (Sci. Agric. Sinica) 37, 1138–1143.

Guan, R., Jiang, S., Xin, R. and Zhang, H.S. (2007) Studies on rapeseed germplasm enhancement by use of cruciferous weed *Descurainia sophia*. *Proceedings of the GCIRC 12th International Rapeseed Congress,* Wuhan, China, 26–30 March 2007, pp. 261–265.

Guan, Z.Q., Chai, T.Y., Zhang, Y.X., Xu, J., Wei, W., *et al*. (2008) Gene manipulation of a heavy metal hyperaccumulator species *Thlaspi caerulescens* L. via Agrobacterium-mediated transformation. *Molecular Biotechnology* 40, 77–86.

Gugel, R.K. and Falk, K.C. (2006) Agronomic and seed quality evaluation of *Camelina sativa* in western Canada. *Canadian Journal of Plant Science* 86, 1047–1058.

Gugel, R.K. and Seguin-Swartz, G. (1997) Introgression of blackleg resistance from *Sinapis alba* into *Brassica napus*. *Brassica 97, Int Soc Hortic Sci Symp Brassicas/10th Crucifer Genetics Workshop*, Rennes, France, 23–27 September 1997, 222 pp.

Gundimeda, H.R., Prakash, S. and Shivanna, K.R. (1992) Intergeneric hybrids between *Enarthrocarpus lyratus*, a wild species, and crop brassicas. *Theoretical and Applied Genetics* 83, 655–662.

Guo, X., Shi, Q., Wang, J., Hou, Y., Wang, Y., *et al*. (2015) Characterization and genome changes of new amphiploids from wheat wide hybridization. *Journal of Genetics and Genomics* 42, 459–461.

Gupta, M., Mason, A.S., Batley, J., Bharti, S., Banga, S., *et al*. (2016) Molecular-cytogenetic characterization of C-genome chromosome substitution lines in *Brassica juncea* (L.) Czern and Coss. *Theoretical and Applied Genetics* 129, 1153–1166.

Gupta, S.B. (1969) Duration of mitotic cycle and regulation of DNA replication in *Nicotiana plumbaginifolia* and a hybrid derivative of *N. tabacum* showing chromosome instability. *Canadian Journal of Genetics and Cytology* 11, 133–142.

Hagimori, M., Nagaoka, M., Kato, N. and Yoshikawa, H. (1992) Production and characterization of somatic hybrids between the Japanese radish and cauliflower. *Theoretical and Applied Genetics* 84, 819–824.

Hall, R.D., Krens, F.A. and Rouwendal, G.J. (1992) DNA radiation damage and asymmetric somatic hybridization: Is UV a potential substitute or supplement to ionising radiation in fusion experiments? *Physiologia Plantarum* 85, 319–324.

Hansen, L.N. (1997). Intertribal somatic hybridization between *Brassica oleracea* L. and *Camelina sativa* (L.) Crantz. *Cruciferae Newsletter* 19, 55–56.

Hansen, L.N. (1998) Intertribal somatic hybridization between rapid cycling *Brassica oleracea* L. and *Camelina sativa* (L.) Crantz. *Euphytica* 104, 173–179.

Hansen, L.N. and Earle, E.D. (1994) Novel flowering and fatty acid characters in rapid cycling *Brassica napus* L. resynthesized by protoplast fusion. *Plant Cell Reports* 14, 151–156.

Hansen, L.N. and Earle, E.D. (1995) Transfer of resistance to *Xanthomonas campestris* pv *campestris* into *Brassica oleracea* L. by protoplast fusion. *Theoretical and Applied Genetics* 91, 1293–1300.

Hanson, B.D., Park, K.W., Mallory-Smith, C.A. and Thill, D.C. (2004) Resistance of *Camelina microcarpa* to acetolactate synthase inhibiting herbicides. *Weed Research* 44, 187–194.

Harberd, D.J. and McArthur, E.D. (1980) Meiotic analysis of some species and genus hybrids in the Brassiceae. In: Tsunoda, S., Hinata, K. and Gomez-Campo, C. (eds) *Brassica Crops and Wild Allies*. Japan Scientific Societies Press, Tokyo, pp. 65–87.

Harlan, J. (1976) Genetic Resources in Wild Relatives of Crops. *Crop Science* 16, 329–333.

Hashem, A., Bowran, D., Piper, T. and Dhammu, H. (2001) Resistance of wild radish (*Raphanus raphanistrum*) to acetolactate synthase-inhibiting herbicides in the Western Australia wheat belt. *Weed Technology* 15, 68–74.

Heap, I.M. (2009) International survey of herbicide-resistant weeds. Available at: http://www.weedscience.org (accessed 21 February 2019).

Heath, D.W. and Earle, E.D. (1997) Synthesis of low linolenic acid rapeseed (*Brassica napus* L.) through protoplast fusion. *Euphytica* 93, 339–343.

Henderson, A.E., Hallett, R.H. and Soroka, J.J. (2004) Prefeeding behavior of the crucifer flea beetle, *Phyllotreta cruciferae*, on host and nonhost crucifers. *Journal of Insect Behavior* 17, 17–39.

Heneen, W.K., Geleta, M., Brismar, K., Xiong, Z., Pires, J.C., *et al*. (2012) Seed colour loci, homoeology and linkage groups of the C genome chromosomes revealed in *Brassica rapa–B. oleracea* monosomic alien addition lines. *Annals of Botany* 109, 1227–1242.

Hermsen, J.T. (1984) Some fundamental considerations on interspecific hybridization. *Iowa State Journal of Research* 58, 461–474.

Heslop-Harrison, J.S. (1999) Aspects of the cell biology of pollination and wide hybridization. In: Cresti, M., Cai, G. and Moscatelli, A. (eds) *Fertilization in Higher Plants (Molecular and Cytological Aspects)*. Springer, Berlin/Heidelberg, Germany, pp. 139–144.

Heyn, F.W. (1976) Transfer of restorer genes from *Raphanus* to cytoplasmic male sterile *Brassica napus*. *Cruciferae Newslette*r 1, 15–16.

Hinata, K. (1979) Studies on a male sterile strain having the *Brassica campestris* nucleus and the *Diplotaxis muralis* cytoplasm. On the breeding procedure and some characteristics of the male sterile strain. *Japanese Journal of Breeding* 29, 305–311.

Hiscock, S.J. and Dickinson, H.G. (1993) Unilateral incompatibility within the Brassicaceae: Further evidence for the involvement of the self-incompatibility (S)-locus. *Theoretical and Applied Genetics* 86, 744–753.

Hoffmann, F. and Adachi, T. (1981) '*Arabidobrassica*': Chromosomal recombination and morphogenesis in asymmetric intergeneric hybrid cells. *Planta* 153(6), 586–593.

Hossain, M.M. and Asahira, T. (1992) Development of heat tolerant somatic hybrids by PEG-mediated protoplasts fusion between *Brassica oleracea* L. and *Brassica campestris* L. *Plant Tissue Culture* 2, 61–69.

Hu, Q., Andersen, S., Dixelius, C. and Hansen, L. (2002a) Production of fertile intergeneric somatic hybrids between *Brassica napus* and *Sinapis arvensis* for the enrichment of the rapeseed gene pool. *Plant Cell Reports* 21, 147–152.

Hu, Q., Hansen, L., Laursen, J., Dixelius, C. and Andersen, S. (2002b) Intergeneric hybrids between *Brassica napus* and *Orychophragmus violaceus* containing traits of agronomic importance for oilseed rape breeding. *Theoretical and Applied Genetics* 105, 834–840.

Hu, Q., Li, Y., Mei, D., Fang, X., Hansen, N.L., *et al*. (2004) Establishment and Identification of cytoplasmic male sterility in *Brassica napus* by intergeneric somatic hybridization. *Zhongguo Nongye Kexue* 37, 333–338.

Hua, L., Wang, D.R., Tan, L., Fu, Y., Liu, F., *et al*. (2015) *LABA1*, a domestication gene associated with long, barbed awns in wild rice. *The Plant Cell* 27, 1875–1888.

Hua, Y.W. and Li, Z.Y. (2006) Genomic *in situ* hybridization analysis of *Brassica napus* × *Orychophragmus violaceus* hybrids and production of *B. napus* aneuploids. *Plant Breeding* 125, 144–149.

Huck, N., Moore, J.M., Federer, M. and Grossniklaus, U. (2003) The *Arabidopsis* mutant feronia disrupts the female gametophytic control of pollen tube reception. *Development* 130, 2149–2159.

Inan, G., Zhang, Q., Li, P., Wang, Z., Cao, Z., *et al*. (2004) Salt cress. A halophyte and cryophyte *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and development of extremophiles. *Plant Physiology* 135, 1718–1737.

Ishikawa, S., Bang, S.W., Kaneko, Y., Matsuzawa, Y. and Robbelen, G. (2003) Production and characterization of intergeneric somatic hybrids between *Moricandia arvensis* and *Brassica oleracea*. *Plant Breeding* 122, 233–238.

Jahier, J., Chevre, A. M., Tanguy, A.M. and Eber, F. (1989) Extraction of disomic addition lines of *Brassica napus–B. nigra*. *Genome* 32, 408–413.

Jain, A., Bhatia, S., Banga, S.S., Prakash, S. and Lakshmikumaran, M. (1994) Potential use of random amplified polymorphic DNA (RAPD) technique to study the genetic diversity in Indian mustard (*Brassica juncea*) and its relationship to heterosis. *Theoretical and Applied Genetics* 88, 116–122.

Janeja, H.S., Banga, S.K., Bhaskar, P.B. and Banga, S.S. (2003) Alloplasmic male sterile *Brassica napus* with *Enarthrocarpus lyratus* cytoplasm: Introgression and molecular mapping of an *E. lyratus* chromosome segment carrying a fertility restoring gene. *Genome* 46, 792–797.

Jarl, C.I. and Bornman, C.H. (1988) Correction of chlorophyll-defective, male-sterile winter oilseed rape (*Brassica napus*) through organelle exchange: Phenotypic evaluation of progeny. *Hereditas* 108, 97–102. DOI: 10.1111/j.1601-5223.1988.tb00687.x.

Jenczewski, E., Eber, F., Grimaud, A., Huet, S., Lucas, M.O., *et al*. (2003) *PrBn*, a major gene controlling homeologous pairing in oilseed rape (*Brassica napus*) haploids. *Genetics* 164, 645–653.

Jeong, B.H., Saga, T., Okayasu, K., Hattori, G., Kaneko, Y., *et al*. (2009) Production and characterization of an amphidiploid line between *Brassica rapa* and a wild relative *Diplotaxis tenuifolia*. *Plant Breeding* 128, 536–537.

Jesske, T., Olberg, B., Schierholt, A. and Becker, H.C. (2013) Resynthesized lines from domesticated and wild *Brassica* taxa and their hybrids with *B. napus* L.: Genetic diversity and hybrid yield. *Theoretical and Applied Genetics* 126, 1053–1065.

Jeuken, M.J., Zhang, N.W., McHale, L.K., Pelgrom, K., Den Boer, E., *et al*. (2009) Rin4 causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. *The Plant Cell* 21, 3368–3378.

Jin, W., Melo, J.R., Nagaki, K., Talbert, P.B., Henikoff, S., *et al*. (2004) Maize centromeres: Organization and functional adaptation in the genetic background of oat. *The Plant Cell* 16, 571–581.

Johnston, S.A. and Hanneman, R.E. (1982) Manipulations of endosperm balance number overcome crossing barriers between diploid *Solanum* species. *Science* 217, 446–448.

Johnston, S.A., Den Nijs, T.P.M., Peloquin, S.J. and Hanneman, R. E. (1980) The significance of genic balance to endosperm development in interspecific crosses. *Theoretical and Applied Genetics* 57, 5–9.

Jugulam, M., McLean, M.D. and Hall, J.C. (2005) Inheritance of picloram and 2, 4-D resistance in wild mustard (*Brassica kaber*). *Weed Science* 53, 417–423.

Jyoti, J.L., Shelton, A.M. and Earle, E.D. (2001) Identifying sources and mechanisms of resistance in crucifers for control of cabbage maggot (Diptera: Anthomyiidae). *Journal of Economic Entomology* 94, 942–949.

Kameya, T. and Hinata, K. (1970) Test-tube fertilization of excised ovules in *Brassica*. *Japanese Journal of Breeding* 20, 253–260.

Kaneko, Y., Matsuzawa, Y. and Sarashima, M. (1987) Breeding of the chromosome addition lines of radish with single kale chromosome. *Japanese Journal of Breeding* 37, 438–452.

Kaneko, Y., Yano, H., Bang, S.W. and Matsuzawa, Y. (2001) Production and characterization of *Raphanus sativus–Brassica rapa* monosomic addition lines. *Plant Breeding* 120, 163–168.

Kaneko, Y., Bang, S.W. and Matsuzawa, Y. (2009) Distant Hybridization. In: Gupta, S.K. (ed.) *Biology and Breeding of Crucifers*. Taylor & Francis Group, New York, pp. 207–247.

Kang, L., Du, X., Zhou, Y., Zhu, B., Ge, X., *et al*. (2014) Development of a complete set of monosomic alien addition lines between *Brassica napus* and *Isatis indigotica* (Chinese woad). *Plant Cell Reports* 33, 1355–1364.

Kang, L., Li, P., Wang, A., Ge, X. and Li, Z. (2017) A novel cytoplasmic male sterility in *Brassica napus* (inap CMS) with carpelloid stamens via protoplast fusion with Chinese woad. *Frontiers in Plant Science* 8, 529. DOI: 10.3389/fpls.2017.00529.

Kang, M.S. and Banga. S.S. (2013) Global agriculture and climate change. In: Kang, M.S. and Banga, S.S. (eds) *Combating Climate Change: An Agricultural Perspective*. CRC Press, Boca Raton, Florida, pp.11–28.

Kao, H.M., Keller, W.A., Gleddie, S. and Brown, G.G. (1992) Synthesis of *Brassica oleracea*/*Brassica napus* somatic hybrid plants with novel organelle DNA compositions. *Theoretical and Applied Genetics* 83, 313–320.

Karpechenko, G. D. (1924) Hybrids of *Raphanus sativus* L. × *Brassica oleracea* L. *Journal of Genetics* 14, 375–396.

Kasha, K.J. and Kao, K.N. (1970) High frequency haploid production in barley (*Hordeum vulgare* L.). *Nature* 225, 874–876.

Kerhoas, C., Knox, R.B. Dumas, C. (1983) Specificity of the callose response in stigmas of *Brassica*. *Annals of Botany* 52, 597–602.

Kift, N.B., Ellis, P.R., Reynolds, K.A., Sime, S. and Pink. D.A.C. (2000) Resistance in wild *Brassica* species to *Brevicoryne brassicae* in the field is not reproduced in the glasshouse, but increases with age. *Proceedings of the 3rd ISHS International Symposium on Brassicas/12th Cruciferae Genetics Workshop*, Wellesbourne, UK, 5–9 September 2000, p. 82.

Kilian, B., Martin, W. and Salamini, F. (2010) Genetic diversity, evolution and domestication of wheat and barley in the Fertile Crescent. In: Glaubrecht, M. (ed.) *Evolution in Action*. Springer, Berlin/Heidelberg, Germany, pp. 137–166.

Kim, N.S., Armstrong, K.C., Fedak, G., Ho, K. and Park, N.I. (2002) A microsatellite sequence from the rice blast fungus (*Magnaporthe grisea*) distinguishes between the centromeres of *Hordeum vulgare* and *H. bulbosum* in hybrid plants. *Genome* 45, 165–174.

Kim, S.Y., Park, B.S., Kwon, S.J., Kim, J., Lim, M.H., *et al*. (2007) Delayed flowering time in *Arabidopsis* and *Brassica rapa* by the overexpression of FLOWERING LOCUS C (FLC) homologs isolated from Chinese cabbage (*Brassica rapa* L. ssp. pekinensis). *Plant Cell Reports* 26, 327–336.

Kinoshita, T. (2007) Reproductive barrier and genomic imprinting in the endosperm of flowering plants. *Genes & Genetic Systems* 82, 177–186.

Kirti, P.B., Prakash, S. and Chopra, V.L. (1991) Interspecific hybridization between *Brassica juncea* and *B. spinescens* through protoplast fusion. *Plant Cell Reports* 9, 639–642.

Kirti, P.B., Narasimhulu, S.B., Prakash, S. and Chopra, V.L. (1992a). Somatic hybridization between *Brassica juncea* and *Moricandia arvensis* by protoplast fusion. *Plant Cell Reports* 11, 318–321.

Kirti, P.B., Narasimhulu, S.B., Prakash, S. and Chopra, V.L. (1992b) Production and characterization of intergeneric somatic hybrids of *Trachystoma ballii* and *Brassica juncea*. *Plant Cell Reports* 11, 90–92.

Kirti, P.B., Narasimhulu, S.B., Mohapatra, T., Prakash, S. and Chopra, V.L. (1993) Correction of chlorophyll deficiency in alloplasmic male sterile *Brassica juncea* through recombination between chloroplast genomes. *Genetics Research* 62, 11–14.

Kirti, P.B., Mohapatra, T., Khanna, H., Prakash, S. and Chopra, V.L. (1995a) *Diplotaxis catholica* + *Brassica juncea* somatic hybrids: Molecular and cytogenetic characterization. *Plant Cell Reports* 14, 593–597.

Kirti, P.B., Mohapatra, T., Baldev, A., Prakash, S. and Chopra, V.L. (1995b) A stable cytoplasmic male-sterile line of *Brassica juncea* carrying restructured organelle genomes from the somatic hybrid *Trachystoma ballii* + *B. juncea*. *Plant Breeding* 114, 434–438.

Kirti, P.B., Banga, S.S., Prakash, S. and Chopra, V.L. (1995c) Transfer of *Ogu* cytoplasmic male sterility to *Brassica juncea* and improvement of the male sterile line through somatic cell fusion. *Theoretical and Applied Genetics* 91, 517–521.

Kirti, P.B., Baldev, A., Gaikwad, K., Bhat, S.R., Kumar, V.D., *et al*. (1997) Introgression of a gene restoring fertility to CMS (*Trachystoma*) *Brassica juncea* and the genetics of restoration. *Plant Breeding* 116, 259–262.

Kirti, P.B., Prakash, S., Gaikwad, K., Kumar, V.D., Bhat, S.R., *et al*. (1998) Chloroplast substitution overcomes leaf chlorosis in a *Moricandia arvensis*-based cytoplasmic male sterile *Brassica juncea*. *Theoretical and Applied Genetics* 97, 1179–1182.

Kiyosue, T., Ohad, N., Yadegari, R., Hannon, M., Dinneny, J., *et al*. (1999) Control of fertilization-independent endosperm development by the MEDEA polycomb gene in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the USA* 96, 4186–4191.

Klewer, A., Scheunemann, R. and Sacristan, M.D. (2003) Incorporation of blackspot resistance from different origins into oilseed rape. *Proceedings of GCIRC 11th International Rapeseed Congress*, Copenhagen, Denmark, 6–10 July 2003, pp. 65–67.

Klimaszewska, K. and Keller, W.A. (1988) Regeneration and characterization of somatic hybrids between *Brassica napus* and *Diplotaxis harra*. *Plant Science* 58, 211–222.

Kmec, P., Weiss, M.J., Milbrath, L.R., Schatz, B.G., Hanzel, J., *et al*. (1998) Growth analysis of *Crambe*. *Crop Science* 38, 108–112.

Kolte, S.J., Bordoloi, D.K. and Awasthi, R.P. (1991) The search for resistance to major diseases of rapeseed and mustard in India. *Proceedings of GCIRC 8th International Rapeseed Congress,* Saskatoon, Canada, 9–11 July 1991, pp. 219–225.

Kondo, K., Yamamoto, M., Matton, D.P., Sato, T., Hirai, M., *et al*. (2002) Cultivated tomato has defects in both S-RNase and HT genes required for stylar function of self-incompatibility. *The Plant Journal* 29, 627–636.

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., *et al*. (2006) An SNP caused loss of seed shattering during rice domestication. *Science* 312, 1392–1396.

Kumar, P., Vasupalli, N., Srinivasan, R. and Bhat, S.R. (2012) An evolutionarily conserved mitochondrial *orf108* is associated with cytoplasmic male sterility in different alloplasmic lines of *Brassica juncea* and induces male sterility in transgenic *Arabidopsis thaliana*. *Journal of Experimental Botany* 63, 2921–2932.

Kumar, S., Atri, C., Sangha, M.K. and Banga, S.S. (2011) Screening of wild crucifers for resistance to mustard aphid, *Lipaphis erysimi* (Kaltenbach) and attempt at introgression of resistance gene (s) from *Brassica fruticulosa* to *Brassica juncea*. *Euphytica* 179, 461–470.

Kumari, P., Bisht, D.S. and Bhat, S.R. (2018) Stable, fertile somatic hybrids between *Sinapis alba* and *Brassica juncea* show resistance to *Alternaria brassicae* and heat stress. *Plant Cell, Tissue and Organ Culture* 133, 77–86.

Lamarck, J.B.A. (1784) *'Chou'. Encyclopédie Methodique Botanique*. I. Paris, France

Lamb, R.J. (1980) Hairs protect pods of mustard (Brassica hirta 'Gisilba') from flea beetle feeding damage. *Canadian Journal of Plant Science* 60, 1439–1440.

Laroche, A., Geng, X.M. and Singh, J. (1992) Differentiation of freezing tolerance and vernalization responses in Cruciferae exposed to a low temperature. *Plant, Cell & Environment* 15, 439–445.

Laurie, D.A. and Bennett, M.D. (1989) The timing of chromosome elimination in hexaploid wheat × maize crosses. *Genome* 32, 953–961.

Lefol, C., Seguin-Swartz, G. and Morrall, R.A.A. (1996) Resistance to *Sclerotinia sclerotiorum* in a weed related to canola. *67th Annual Meeting, Canadian Phytopathological Society*. Saskatoon, Canada, 22–26 June 1996.

Lehtila, K. and Strauss, S.Y. (1999) Effects of foliar herbivory on male and female reproductive traits of wild radish, *Raphanus raphanistrum*. *Ecology* 80, 116–124.

Leino, M., Teixeira, R., Landgren, M. and Glimelius, K. (2003) *Brassica napus* lines with rearranged *Arabidopsis* mitochondria display CMS and a range of developmental aberrations. *Theoretical and Applied Genetics* 106, 1156–1163.

Leino, M., Thyselius, S., Landgren, M. and Glimelius, K. (2004) *Arabidopsis thaliana* chromosome III restores fertility in a cytoplasmic male-sterile *Brassica napus* line with *A. thaliana* mitochondrial DNA. *Theoretical and Applied Genetics* 109, 272–279.

Lelivelt, C.L.C. and Krens, F.A. (1992) Transfer of resistance to the beet cyst nematode (*Heterodera schachtii* Schm.) into the *Brassica napus* L. gene pool through intergeneric somatic hybridization with *Raphanus sativus* L. *Theoretical and Applied Genetics* 83, 887–894.

Lelivelt, C.L.C., Leunissen, E.H.M., Frederiks, H.J., Helsper, J.P.F.G. and Krens, F.A. (1993) Transfer of resistance to the beet cyst nematode (*Heterodera schachtii* Schm.) from *Sinapis alba* L. (white mustard) to the *Brassica napus* L. gene pool by means of sexual and somatic hybridization. *Theoretical and Applied Genetics* 85, 688–696.

Li, A., Jiang, J., Zhang, Y., Snowdon, R.J., Liang, G., *et al*. (2012) Molecular and cytological characterization of introgression lines in yellow seed derived from somatic hybrids between *Brassica napus* and *Sinapis alba*. *Molecular Breeding* 29(1), 209–219.

Li, H., Barbetti, M.J. and Sivasithamparam, K. (2005b) Hazard from reliance on cruciferous hosts as sources of major gene-based resistance for managing blackleg (*Leptosphaeria maculans*) disease. *Field Crops Research* 91, 185–198.

Li, P., Kang, L., Wang, A., Cui, C., Jiang, L., Guo, S., Ge, X. and Li, Z. (2019) Development of a fertility restorer for *inap* CMS (*Isatis indigotica*) *Brassica napus* through genetic introgression of one alien addition. *Frontiers in Plant Science* 10, 257. DOI: 10.3389/fpls.2019.00257.

Li, X.F., Xuan, S.X., Wang, J.L., Zhang, S.L., Wang, Y.H., *et al*. (2013) Generation and identification of *Brassica alboglabra–Brassica campestris* monosomic alien addition lines. *Genome* 56, 171–177.

Li, Z. and Heneen, W.K. (1999) Production and cytogenetics of intergeneric hybrids between the three cultivated *Brassica* diploids and *Orychophragmus violaceus*. *Theoretical and Applied Genetics* 99, 694–704.

Li, Z., Liu, H.L. and Luo, P. (1995) Production and cytogenetics of intergeneric hybrids between *Brassica napus* and *Orychophragmus violaceus*. *Theoretical and Applied Genetics* 91, 131–136.

Li, Z., Cartagena, J. and Fukui, K. (2005a) Simultaneous detection of 5S and 45S rRNA genes in *Orychophragmus violaceus* by double fluorescence *in situ* hybridization. *Cytologia* 70, 459–466.

Li, Z., Wu, J.G., Liu, Y., Liu, H.L. and Heneen, W.K. (1998a) Production and cytogenetics of the intergeneric hybrids *Brassica juncea* × *Orychophragmus violaceus* and *B. carinata* × *O. violaceus*. *Theoretical and Applied Genetics* 96, 251–265.

Li, Z.Y., Liu, H.L. and Heneen, W.K. (1996) Meiotic behaviour in intergeneric hybrids between *Brassica napus* and *Orychophragmus violaceus*. *Hereditas* 125, 69–75.

Li, Z.Y., Liang, X.M., Wu, J.G. and Heneen, W.K. (1998b) Morphology and cytogenetics of F_3 progenies from intergeneric hybrids between *Brassica juncea* and *Ovychophragmus violaceus*. *Hereditas* 129, 143–150.

Li, Z.Y., Ceccarelli, M., Minelli, S., Contento, A., Liu, Y., *et al*. (2003) Genomic *in situ* hybridization analysis of intergeneric hybrids between *Brassica* species and *Orychophragmus violaceus* and detection of rDNA loci in *O. violaceus*. *Proceedings of GCIRC 11th International Rapeseed Congress*, Copenhagen, Denmark, 6–10 July 2003, pp. 6–10.

Lian, Y.J., Lin, G.Z., Zhao, X.M. and Lim, H.T. (2011) Production and genetic characterization of somatic hybrids between leaf mustard (*Brassica juncea*) and broccoli (*Brassica oleracea*). *In Vitro Cellular & Developmental Biology – Plant* 47, 289–296.

Linde-Laursen, I.B. and von Bothmer, R. (1999) Orderly arrangement of the chromosomes within barley genomes of chromosome-eliminating *Hordeum lechleri* × barley hybrids. *Genome* 42, 225–236.

Linnaeus, C. (1753) *Species Plantarum II*. Laurentius Salvius, Stockholm, Sweden.

Liu, J.H., Dixelius, C., Eriksson, I. and Glimelius, K. (1995) *Brassica napus* (+) *B. tournefortii*, a somatic hybrid containing traits of agronomic importance for rapeseed breeding. *Plant Science* 109, 75–86.

Liu, J.H., Landgren, M. and Glimelius, K. (1996) Transfer of the *Brassica tournefortii* cytoplasm to *B. napus* for the production of cytoplasmic male sterile *B. napus*. *Physiologia Plantarum* 96, 123–129.

Liu, R., Qian, W. and Meng, J. (2002) Association of RFLP markers and biomass heterosis in trigenomic hybrids of oilseed rape (*Brassica napus* × *B. campestris*). *Theoretical and Applied Genetics* 105, 1050–1057.

Liu, Z., Adamczyk, K., Manzanares-Dauleux, M., Eber, F., Lucas, M.O., *et al*. (2006) Mapping *PrBn* and other quantitative trait loci responsible for the control of homeologous chromosome pairing in oilseed rape (*Brassica napus* L.) haploids. *Genetics* 174, 1583–1596.

Lobell, D.B. and Gourdji, S.M. (2012) The influence of climate change on global crop productivity. *Plant Physiology* 160, 1686–1697.

Lu, J.H., Liu, S.S. and Shelton, A.M. (2004) Laboratory evaluations of a wild crucifer *Barbarea vulgaris* as a management tool for the diamondback moth *Plutella xylostella* (Lepidoptera: Plutellidae). *Bulletin of Entomological Research* 94, 509–516.

Luo, P., Lan, Z.Q., Huang, J. and Li, Z.Y. (1991) Study on valuable plant resource *Orychophragmus violaceus* (L.) OE Schulz. *Journal of Natural Resources* 6, 206–210.

Luo, P., Lan, Z.Q. and Li, Z.Y. (1994) *Orychophragmus violaceus*, a potential edible-oil crop. *Plant Breeding* 113, 83–85.

Ma, N. and Li, Z.Y. (2007) Development of novel *Brassica napus* lines with canola quality and higher levels of oleic and linoleic acids derived from intergeneric hybrids between *B. napus* and *Orychophragmus violaceus*. *Euphytica* 157, 231–238.

Ma, N., Li, Z.Y., Cartagena, J.A. and Fukui, K. (2006) GISH and AFLP analyses of novel *Brassica napus* lines derived from one hybrid between *B. napus* and *Orychophragmus violaceus*. *Plant Cell Reports* 25, 1089–1093.

Malik, M., Vyas, P., Rangaswamy, N.S. and Shivanna, K.R. (1999) Development of two new cytoplasmic male-sterile lines in *Brassica juncea* through wide hybridization. *Plant Breeding* 118, 75–78.

Mamula, D., Juretic, N. and Horvath, J. (1997) Susceptibility of host plants to belladonna mottle and turnip yellow mosaic tymoviruses: Multiplication and distribution. *Acta Phytopathologica et Entomologica Hungarica* 32, 289–298.

Marvin, H.J., Mastebroek, H.D., Becu, D.M. and Janssens, R.J. (2000) Investigation into the prospects of five novel oilseed crops within Europe. *Outlook on Agriculture* 29, 47–53.

Mathews, S., Singhal, R.S. and Kulkarni, P.R. (1993) Some physicochemical characteristics of *Lepidium sativum* (haliv) seeds. *Nahrung* 37, 69–71. DOI: 10.1002/food.19930370113.

Mathias, R. (1985) Transfer of cytoplasmic male sterility from brown mustard (*Brassica juncea* L. Coss.) into rapeseed (*Brassica napus* L.). *Zeitschrift fuer Pflanzenzuechtung* 95, 371–374.

Matsuzawa, Y., Kaneko, Y. and Bang, S. (1996) Prospects of the wide cross for genetics and plant breeding in Brassiceae. *Bulletin of the College of Agriculture, Utsunomiya University* 16, 5–10.

Matsuzawa, Y., Minami, T., Bang, S.W. and Kaneko, Y. (1997) A new *Brassicoraphanus* (2n = 36); the true-breeding amphidiploid line of *Brassica oxyrrhina* Coss. (2n = 18) × *Raphanus sativus* L. (2n = 18). *Bulletin of the College of Agriculture*, *Utsunomiya University* 16, 1–7.

Matsuzawa, Y., Mekiyanon, S., Kaneko, Y., Bang, S.W., Wakui, K., *et al*. (1999) Male sterility in alloplasmic *Brassica rapa* L. carrying *Eruca sativa* cytoplasm. *Plant Breeding* 118, 82–84.

Mayr, E. (1947) Ecological factors in speciation. *Evolution* 1, 263–288.
McGrath, J.M. and Quiros, C.F. (1990) Generation of alien chromosome addition lines from synthetic *Brassica napus*: Morphology, cytology, fertility, and chromosome transmission. *Genome* 33, 374–383.
McLellan, M.S., Olesen, P. and Power, J.B. (1988) Towards the introduction of cytoplasmic male sterility (CMS) into *Brassica napus* through protoplast fusion. In: Puite, K.J., Dons, J.J.M., Huizing, H.J., Kool, A.J., Koornneef, M., *et al* (eds) *Progress in Plant Protoplast Research. Current Plant Science and Biotechnology in Agriculture*. Vol. 7, Springer, Dordrecht, The Netherlands, pp. 187–188.
Mei, D., Li, Y. and Hu, Q. (2003) Study of male sterile line derived from intergeneric hybrids of *Brassica napus* + *Orychophragmus violaceus* and *B. napus* + *Sinapis arvensis*. *Chinese Journal of Oil Crop Sciences* 25, 72–75.
Menczel, L., Morgan, A., Brown, S. and Maliga, P. (1987) Fusion-mediated combination of Ogura-type cytoplasmic male sterility with *Brassica napus* plastids using X-irradiated CMS protoplasts. *Plant Cell Reports* 6, 98–101.
Meng, J.L., Yan, Z., Tian, Z., Huang, R. and Huang, B. (1999) Somatic hybrids between *Moricandia nitens* and three *Brassica* species. *Proceedings of the 10th International Rapeseed Congress*, Canberra, Australia, 26–29 September 1999, pp. 4–8.
Meyer, R.S., DuVal, A.E. and Jensen, H.R. (2012) Patterns and processes in crop domestication: An historical review and quantitative analysis of 203 global food crops. *New Phytologist* 196, 29–48.
Miller, A.J. and Gross, B.L. (2011) From forest to field: Perennial fruit crop domestication. *American Journal of Botany* 98, 1389–1414.
Mithen, R.F. and Herron, C. (1991) Transfer of disease resistance to oilseed rape from wild *Brassica* species. *Proceedings of the 8th GCIRC International Rapeseed Congress*, Saskatoon, Canada, 9–11 July 1991, pp. 244–249.
Mithen, R.F. and Magrath, R. (1992) Glucosinolates and resistance to *Leptosphaeria maculans* in wild and cultivated *Brassica* species. *Plant Breeding* 108, 60–68.
Mithen, R.F., Lewis, B.G., Heaney, R.K. and Fenwick, G.R. (1987) Resistance of leaves of *Brassica* species to *Leptosphaeria maculans*. *Transactions of the British Mycological Society* 88, 525–531.
Mithila, J. and Hall, J.C. (2007) Production of an auxinic herbicide-resistant microspore-derived doubled haploid wild mustard (*Sinapis arvensis* L.) plant. *Crop Protection* 26, 357–362.
Mithila, J. and Hall, J.C. (2013) Transfer of auxinic herbicide resistance from *Brassica kaber* to *Brassica juncea* and *Brassica rapa* through embryo rescue. *In Vitro Cellular & Developmental Biology – Plant* 49, 461–467.
Mizushima, U. (1950) Karyogenetic studies of species and genus hybrids in the tribe Brassiceae of Cruciferae. *Tohoku Journal of Agricultural Research* 1, 1–14.
Mizushima, U. (1968) Phylogenetic studies on some wild *Brassica* species. *Tohoku Journal of Agricultural Research* 19, 83–99.
Mochida, K., Tsujimoto, H. and Sasakuma, T. (2004) Confocal analysis of chromosome behavior in wheat × maize zygotes. *Genome* 47, 199–205.
Mohanty, A., Chrungu, B., Verma, N. and Shivanna, K.R. (2009) Broadening the genetic base of crop brassicas by production of new intergeneric hybrid. *Czech Journal of Genetics and Plant Breeding* 45, 117–122.
Mohapatra, T., Kirti, P.B., Kumar, V.D., Prakash, S. and Chopra, V.L. (1998) Random chloroplast segregation and mitochondrial genome recombination in somatic hybrid plants of *Diplotaxis catholica* + *Brassica juncea*. *Plant Cell Reports* 17, 814–818.
Muller, H. (1942) Isolating mechanisms, evolution, and temperature. *Biology Symposia* 6, 71–125.
Murase, K., Shiba, H., Iwano, M., Che, F.S., Watanabe, M., *et al*. (2004) A membrane-anchored protein kinase involved in *Brassica* self-incompatibility signaling. *Science* 303, 1516–1519.
Murfett, J., Strabala, T.J., Zurek, D.M., Mou, B., Beecher, B., *et al*. (1996) S RNase and interspecific pollen rejection in the genus *Nicotiana*: Multiple pollen-rejection pathways contribute to unilateral incompatibility between self-incompatible and self-compatible species. *The Plant Cell* 8, 943–958.
Nagaharu, U. (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany* 7, 389–452.
Namai, H. (1987) Inducing cytogenetical alterations by means of interspecific and intergeneric hybridization in *Brassica* crops. *Gamma Field Symposia* 26, 41–89.
Narasimhulu, S.B., Kirti, P.B., Bhatt, S.R., Prakash, S. and Chopra, V.L. (1994) Intergeneric protoplast fusion between *Brassica carinata* and *Camelina sativa*. *Plant Cell Reports* 13, 657–660.

Naumova, T.N., van der Laak, J., Osadtchiy, J., Matzk, F., Kravtchenko, A., *et al*. (2001) Reproductive development in apomictic populations of *Arabis holboellii* (Brassicaceae). *Sexual Plant Reproduction* 14, 195–200.

Navabi, Z.K., Parkin, I.A., Pires, J.C., Xiong, Z., Thiagarajah, M.R., *et al*. (2010) Introgression of B-genome chromosomes in a doubled haploid population of *Brassica napus* × *B. carinata*. *Genome* 53, 619–629. DOI: 10.1139/G10-039.

Navratilova, B., Buzek, J., Siroky, J. and Havranek, P. (1997) Construction of intergeneric somatic hybrids between *Brassica oleracea* and *Armoracia rusticana*. *Biologia Plantarum* 39, 531–541.

Nitovska, I.A., Shakhovsky, A.M., Cherep, N.N., Gorodenska, M.M., Kuchuk, N.V., *et al*. (2006) Construction of the cybrid transplastomic *Brassica napus* plants containing *Lesquerella fendleri* chloroplasts. *Tsitologiya i Genetika* 40, 3–11.

Ogura, H. (1968) Studies on the new male-sterility in Japanese radish, with special reference to the utilization of this sterility towerds the practical raising of hybrid seeds. *Memoirs of the Faculty of Agriculture, Kagoshima University* 6, 39–78.

Oikarinen, S. and Ryoppy, P.H. (1992) Somatic hybridization of *Brassica campestris* and *Barbarea* species. *Proceedings of XIIIth Eucarpia Congress: Reproductive Biology and Plant Breeding*, Angers, France, 6–11 July, pp. 261–262.

Okamoto, S., Odashima, M., Fujimoto, R., Sato, Y., Kitashiba, H., *et al*. (2007) Self-compatibility in *Brassica napus* is caused by independent mutations in S-locus genes. *The Plant Journal* 50, 391–400.

O'Neill, C.M., Murata, T., Morgan, C.L. and Mathias, R.J. (1996) Expression of the C3-C4 intermediate character in somatic hybrids between *Brassica napus* and the C_3-C_4 species *Moricandia arvensis*. *Theoretical and Applied Genetics* 93, 1234–1241.

Osborn, T.C. (2004) The contribution of polyploidy to variation in *Brassica* species. *Physiologia Plantarum* 121, 531–536. DOI: 10.1111/j.1399-3054.2004.00360.x.

Ovcharenko, O.O., Komarnyts-kyi, I.K., Cherep, M.M., Hleba, I. and Kuchuk, M.V. (2004) Obtaining of intertribal *Brassica juncea* + *Arabidopsis thaliana* somatic hybrids and study of transgenic trait behaviour. *Tsitologiya i Genetika* 38, 3–8.

Ovcharenko, O., Momot, V., Cherep, N., Sheludko, Y., Komarnitsky, I., *et al*. (2011) Transfer of transformed *Lesquerella fendleri* (Gray) Wats. chloroplasts into *Orychophragmus violaceus* (L.) OE Schulz by protoplast fusion. *Plant Cell,Tissue and Organ Culture*, 105, 21–27.

Pathania, A., Bhat, S.R., Kumar, V.D., Kirti, P.B., Prakash, S., *et al*. (2003) Cytoplasmic male sterility in alloplasmic *Brassica juncea* carrying *Diplotaxis catholica* cytoplasm: Molecular characterization and genetics of fertility restoration. *Theoretical and Applied Genetics* 107, 455–461.

Pathania, A., Kumar, R., Kumar, V.D., Asutosh, K.K., Dwivedi, Kirti, P.B., *et al*. (2007) A duplication of coxI gene is associated with CMS (*Diplotaxis catholica*) *Brassica juncea* derived from somatic hybridization with *Diplotaxis catholica*. *Journal of Genetics* 86, 93–101.

Pattison, A.B., Versteeg, C., Akiew, S. and Kirkegaard, J. (2006) Resistance of Brassicaceae plants to root-knot nematode (*Meloidogyne* spp.) in northern Australia. *International Journal of Pest Management* 52, 53–62.

Paulmann, W. and Robbelen, G. (1988) Effective transfer of cytoplasmic male sterility from radish (*Raphanus sativus* L.) to rape (*Brassica napus* L.). *Plant Breeding* 100, 299–309.

Paulose, B., Zulfiqar, A. and Parkash, O. (2007) Isolation and characterization of arsenic induced genes from *Crambe abyssinica*. *71st annual Meeting of the Northeast Section of the American Society of Plant Biology – Fueling the Future through Plant Biology*, Syracuse, New York, USA, 1–2 June 2007, pp. 1–2.

Pedras, M.S.C. and Adio, A.M. (2008) Phytoalexins and phytoanticipins from the wild crucifers *Thellungiella halophila* and *Arabidopsis thaliana*: Rapalexin A, wasalexins and camalexin. *Phytochemistry* 69, 889–893.

Pedras, M.S.C., Chumala, P.B. and Suchy, M. (2003) Phytoalexins from *Thlaspi arvense*, a wild crucifer resistant to virulent *Leptosphaeria maculans*: Structures, syntheses and antifungal activity. *Phytochemistry* 64, 949–956.

Pellan-Delourme, R. and Renard, M. (1987) Identification of maintainer genes in *Brassica napus* L. for the male-sterility-inducing cytoplasm of *Diplotaxis muralis* L. *Plant Breeding* 99, 89–97.

Pelletier, G., Primard, C., Vedel, F., Chetrit, P., Remy, R., *et al*. (1983) Intergeneric cytoplasmic hybridization in Cruciferae by protoplast fusion. *Molecular and General Genetics* 191, 244–250.

Pelletier, G., Ferault, M., Lancelin, D. and Boulidard, L. (1989) *Brassica oleracea* cybrids and their potential for hybrid seed production. *Vortraege fuer Pflanzenzuechtung* 7, 15.

Peterka, H., Budahn, H., Schrader, O., Ahne, R. and Schütze, W. (2004) Transfer of resistance against the beet cyst nematode from radish (*Raphanus sativus*) to rape (*Brassica napus*) by monosomic chromosome addition. *Theoretical and Applied Genetics* 109, 30–41.

Pollard, A.J. and Baker, A.J. (1996) Quantitative genetics of zinc hyperaccumulation in *Thlaspi caerulescens*. *The New Phytologist* 132, 113–118.

Pollard, A.J. and Baker, A.J. (1997) Deterrence of herbivory by zinc hyperaccumulation in *Thlaspi caerulescens* (Brassicaceae). *The New Phytologist* 135, 655–658.

Pradhan, A.K., Mukhopadhyay, A. and Pental, D. (1991) Identification of the putative cytoplasmic donor of a CMS system in *Brassica juncea*. *Plant Breeding* 106, 204–208.

Pradhan, A.K., Prakash, S., Mukhopadhyay, A. and Pental, D. (1992) Phytogeny of *Brassica* and allied genera based on variation in chloroplast and mitochondrial DNA patterns: Molecular and taxonomic classifications are incongruous. *Theoretical and Applied Genetics* 85, 331–340.

Prakash, S. (2010) *Brassica* cytogenetics – a historical journey and my personal reminiscence. *Chinese Journal of Oil Crops* 32, 163–172.

Prakash, S. and Chopra, V.L. (1988) Synthesis of alloplasmic *Brassica campestris* and induction of cytoplasmic male sterility. *Plant Breeding* 101, 253–255.

Prakash, S. and Chopra, V.L. (1990) Male sterility caused by cytoplasm of *Brassica oxyrrhina* in *B. campestris* and B. juncea. *Theoretical and Applied Genetics* 79, 285–287.

Prakash, S. and Bhat, S.R. (2007) Contribution of wild crucifers in *Brassica* improvement: Past accomplishment and future perspectives. *Proceedings of the 12th GCIRC International Rapeseed Congress*, Wuhan, China, 26–30 March 2007, pp. 213–215.

Prakash, S., Kirti, P.B., Bhat, S.R., Gaikwad, K., Kumar, V.D., *et al*. (1998) A *Moricandia arvensis*–based cytoplasmic male sterility and fertility restoration system in *Brassica juncea*. *Theoretical and Applied Genetics* 97, 488–492.

Prakash, S., Ahuja, I., Upreti, H.C., Kumar, V.D., Bhat, S.R., *et al*. (2001) Expression of male sterility in alloplasmic *Brassica juncea* with *Erucastrum canariense* cytoplasm and the development of a fertility restoration system. *Plant Breeding* 120, 479–482.

Prescott-Allen, C. and Prescott-Allen, R. (1986) *The first resource: Wild species in the North American economy*. Yale University Press, New Haven, Connecticut.

Primard, C., Vedel, F., Mathieu, C., Pelletier, G. and Chevre, A.M. (1988) Interspecific somatic hybridization between *Brassica napus* and *Brassica hirta* (*Sinapis alba* L.). T*heoretical and Applied Genetics* 75(4), 546–552.

Primard, C., Poupard, J.P., Horvais, R., Eber, F., Pelletier, G., *et al*. (2005) A new recombined double low restorer line for the *Ogu-INRA* cms in rapeseed (*Brassica napus* L.). *Theoretical and Applied Genetics* 111, 736–746.

Qian, W., Liu, R. and Meng, J. (2003) Genetic effects on biomass yield in interspecific hybrids between *Brassica napus* and *B. rapa*. *Euphytica* 134, 9–15.

Quiros, C.F., Ochoa, O., Kianian, S.F. and Douches, D. (1987) Analysis of the *Brassica oleracea* genome by the generation of *B. campestris-oleracea* chromosome addition lines: Characterization by isozymes and rDNA genes. *Theoretical and Applied Genetics* 74, 758–766.

Raman, R., Qiu, Y., Coombes, N., Song, J., Kilian, A., *et al*. (2017) Molecular diversity analysis and genetic mapping of pod shatter resistance loci in *Brassica carinata* L. *Frontiers in Plant Science* 8, 1765. DOI: 10.3389/fpls.2017.01765.

Ramsey, A.D. and Ellis, P.R. (1994) Resistance in wild brassicas to the cabbage whitefly, *Aleyrodes proletella*. *ISHS Symposium on Brassicas, 9th Crucifer Genetics Workshop*, Lisbon, Portugal, 15–18 November 1994, pp. 507–514.

Rana, J.S., Khokhar, K.S. and Singh, H. (1995) Relative susceptibility of *Brassica* species to mustard aphid, *Lipaphis erysimi* (Kalt.). *Journal of Insect Science* 8, 96–97.

Rana, K., Atri, C., Gupta, M., Akhatar, J., Sandhu, P.S., *et al*. (2017) Mapping resistance responses to *Sclerotinia* infestation in introgression lines of *Brassica juncea* carrying genomic segments from wild Brassicaceae *B. fruticulosa*. *Scientific Reports* 7, 5904. DOI: 10.1038/s41598-017-05992-9.

Rao, G.U. and Shivanna, K.R. (1996) Development of a new alloplasmic CMS *Brassica napus* in the cytoplasmic background of *Diplotaxis siifolia*. *Cruciferae Newsletter* 18, 68–69.

Rao, G.U., Batra-Sarup, V., Prakash, S. and Shivanna, K.R. (1994) Development of a new cytoplasmic male-sterility system in *Brassica juncea* through wide hybridization. *Plant Breeding* 112, 171–174.

Rao, G.U., Lakshmikumaran, M. and Shivanna, K.R. (1996) Production of hybrids, amphiploids and backcross progenies between a cold-tolerant wild species, *Erucastrum abyssinicum* and crop Brassicas. *Theoretical and Applied Genetics* 92, 786–790.

Rashid, M.H., Zou, Z. and Fernando, W.D. (2018) Development of molecular markers linked to the *Leptosphaeria maculans* resistance gene *Rlm6* and inheritance of SCAR and CAPS markers in *Brassica napus* × *Brassica juncea* interspecific hybrids. *Plant Breeding* 137, 402–411.

Rawat, D.S. and Anand, I.J. (1979) Male sterility in Indian mustard. *Indian Journal of Genetics and Plant Breeding* 39, 412–414.

Rehn, F., Arbeiter, A. and Siemens, J. (2004) The gene *RPB1* confers resistance of *Arabidopsis thaliana* to the obligate biotrophic parasite *Plasmodiophora brassicae*. *Proceedings of the 4th ISHS Symposium Brassicas/14th Crucifer Genetics Workshop*, Daejeon, South Korea. 24–28 October 2004, p. 137.

Ren, J.P., Dickson, M.H. and Earle, E.D. (2000) Improved resistance to bacterial soft rot by protoplast fusion between *Brassica rapa* and *B. oleracea*. *Theoretical and Applied Genetics* 100, 810–819.

Renwick, J.A.A. (2002) The chemical world of crucivores: Lures, treats and traps. In: Nielsen J.K., Kjaer C. and Schoonhoven L.M. (eds) *Proceedings of the 11th International Symposium on Insect–Plant Relationships*. Series Entomologica, Springer, Dordrecht, The Netherlands, pp. 35–42. DOI: 10.1007/978-94-017-2776-1_4.

Riera-Lizarazu, O., Rines, H.W. and Phillips, R.L. (1996) Cytological and molecular characterization of oat × maize partial hybrids. *Theoretical and Applied Genetics* 93, 123–135.

Riley, R. and Chapman, V. (1958) Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature* 182, 713–715.

Robinson, B., Duwig, C., Bolan, N., Kannathasan, M. and Saravanan, A. (2003) Uptake of arsenic by New Zealand watercress (*Lepidium sativum*). *Science of the Total Environment* 301, 67–73.

Rotman, N., Rozier, F., Boavida, L., Dumas, C., Berger, F., *et al*. (2003) Female control of male gamete delivery during fertilization in *Arabidopsis thaliana*. *Current Biology* 13, 432–436.

Roux, F., Matejicek, A. and Reboud, X. (2005a) Response of *Arabidopsis thaliana* to 22 ALS inhibitors: Baseline toxicity and cross-resistance of csr1-1 and csr1-2 resistant mutants. *Weed Research* 45, 220–227.

Roux, F., Matejicek, A., Gasquez, J. and Reboud, X. (2005b) Dominance variation across six herbicides of the *Arabidopsis thaliana* csr1-1 and csr1-2 resistance alleles. *Pest Management Science* 61, 1089–1095.

Roy, N.N. (1984) Interspecific transfer of *Brassica juncea*-type high blackleg resistance to *Brassica napus*. *Euphytica* 33, 295–303.

Ryschka, U., Schumann, G., Klocke, E., Scholze, P. and Neumann, M. (1996) Somatic hybridization in brassicaceae. *Acta Horticulturae* 407, 201–208. DOI: 10.17660/ActaHortic.1996.407.24

Ryschka, U., Schumann, G., Klocke, E., Scholze, P. and Kramer, R. (1999) Somatic cell hybridization for transfer of disease resistance in *Brassica*. In: Altman, A., Ziv, M. and Izhar, S. (eds) *Plant Biotechnology and In Vitro Biology in the 21st Century*. Springer, Dordrecht, The Netherlands, pp. 205–208.

Saal, B., Brun, H., Glais, I. and Struss, D. (2004) Identification of a *Brassica juncea*-derived recessive gene conferring resistance to *Leptosphaeria maculans* in oilseed rape. *Plant Breeding* 123, 505–511.

Sageret, M. (1826) Considerations surla production des variants et des varieties en general, et sur celled de la famille de Cucurbitacee s en particulier. *Annales des sciences naturelles* 8, 94–314.

Sakai, T. and Imamura, J. (1990) Intergeneric transfer of cytoplasmic male sterility between *Raphanus sativus* (cms line) and *Brassica napus* through cytoplast-protoplast fusion. *Theoretical and Applied Genetics* 80, 421–427.

Sakai, T., Liu, H.J., Iwabuchi, M., Kohno-Murase, J. and Imamura, J. (1996) Introduction of a gene from fertility restored radish (*Raphanus sativus*) into *Brassica napus* by fusion of X-irradiated protoplasts from a radish restorer line and iodacetoamide-treated protoplasts from a cytoplasmic male-sterile cybrid of *B. napus*. *Theoretical and Applied Genetics* 93, 373–379.

Sakhno, L.A., Komarnitsky, I.K., Cherep, N.N. and Kuchuk, N.V. (2007) Phosphinotricin-resistant somatic hybrids *Brassica napus* + *Orychophragmus violaceus*. *Tsitologiya i Genetika* 41, 3–8.

Salisbury, P.A. (1989) Potential utilization of wild crucifer germplasm in oilseed *Brassica* breeding. *Proceedings of 7th Australian Rapeseed Agronomists and Breeders Workshop*, Toowoomba, Queensland, Australia, 12–17 September 1989, pp. 51–53.

Sampath, A. (2009) Chemical characterization of *camelina* seed oil. PhD dissertation, Rutgers University, New Brunswick, New Jersey, USA.

Sarikamis, G., Marquez, J., MacCormack, R., Bennett, R. N., Roberts, J. and Mithen, R. (2006) High glucosinolate broccoli: A delivery system for sulforaphane. *Molecular Breeding* 18, 219–228.

Sarla, N. (1988) Overcoming interspecific incompatibility in the cross *Brassica campestris* ssp. *japonica* × *Brassica oleracea* var. *botrytis* using irradiated mentor pollen. *Biologia Plantarum* 30, 384–386.

Scholze, P., Kramer, R., Ryschka, U., Klocke, E. and Schumann, G. (2010) Somatic hybrids of vegetable brassicas as source for new resistances to fungal and virus diseases. *Euphytica* 176, 1–14.

Schroder-Pontoppidan, M., Skarzhinskaya, M., Dixelius, C., Stymne, S. and Glimelius, K. (1999) Very long chain and hydroxylated fatty acids in offspring of somatic hybrids between *Brassica napus* and *Lesquerella fendleri*. *Theoretical and Applied Genetics* 99, 108–114.

Schwarzacher, T., Anamthawat-Jonsson, K., Harrison, G.E., Islam, A.K.M.R., Jia, J.Z., *et al*. (1992) Genomic *in situ* hybridization to identify alien chromosomes and chromosome segments in wheat. *Theoretical and Applied Genetics* 84, 778–786.

Schwarzacher-Robinson, T., Finch, R.A., Smith, J.B. and Bennett, M.D. (1987) Genotypic control of centromere positions of parental genomes in *Hordeum* × *Secale* hybrid metaphases. *Journal of Cell Science* 87, 291–304.

Sears, E.R. (1976) Genetic control of chromosome pairing in wheat. *Annual Review of Genetics* 10, 31–51.

Sharma, B.B., Kalia, P., Singh, D. and Sharma, T.R. (2017) Introgression of black rot resistance from *Brassica carinata* to cauliflower (*Brassica oleracea* botrytis group) through embryo rescue. *Frontiers in Plant Science* 8, 1255. DOI: 10.3389/fpls.2017.01255.

Sharma, G., Kumar, V.D., Haque, A., Bhat, S.R., Prakash, S. *et al*. (2002) *Brassica* coenospecies: A rich reservoir for genetic resistance to leaf spot caused by *Alternaria brassicae*. *Euphytica* 125, 411–417.

Sharma, H.C. (1995) How wide can a wide cross be? *Euphytica* 82, 43–64.

Sharma, N., Cram, D., Huebert, T., Zhou, N. and Parkin, I.A. (2007) Exploiting the wild crucifer *Thlaspi arvense* to identify conserved and novel genes expressed during a plant's response to cold stress. *Plant Molecular Biology* 63, 171–184.

Sharma, T.R. and Singh, B.M. (1992) Transfer of resistance to *Alternaria brassicae* in *Brassica juncea* through interspecific hybridization among *Brassicas*. *Journal of Genetics and Breeding* 46, 373–378.

Shea, D.J., Tomaru, Y., Itabashi, E., Nakamura, Y., Miyazaki, T., *et al*. (2018) The production and characterization of a BoFLC2 introgressed *Brassica rapa* by repeated backcrossing to an F_1. *Breeding Science* 68, 316–325. DOI: 10.1270/jsbbs.17115.

Shen, Y., Xiang, Y., Xu, E., Ge, X. and Li, Z. (2018) Major co-localized QTL for plant height, branch initiation height, stem diameter, and flowering time in an alien introgression derived *Brassica napus* DH population. *Frontiers in Plant Science* 9, 390. DOI: 10.3389/fpls.2018.00390.

Shimizu, S., Kanazawa, K. and Kobayashi, T. (1962) Studies on the breeding of Chinese cabbage for resistance to soft rot. Part III. The breeding of the resistant variety 'Hiratsuka No. 1' by interspecific crossing. *Bulletin of the Horticultural Research Station*. *Series A, Japan* 1, 157–174.

Shinada, T., Kikuchi, Y., Fujimoto, R. and Kishitani, S. (2006) An alloplasmic male-sterile line of *Brassica oleracea* harboring the mitochondria from *Diplotaxis muralis* expresses a novel chimeric open reading frame, orf72. *Plant and Cell Physiology* 47, 549–553.

Shivanna, K.R. and Johri, B.M. (1985) *The Angiosperm Pollen: Structure and Function*. Wiley Eastern Limited, New Delhi.

Shrestha, S.K., Mathur, S.B. and Munk, L. (2000) *Alternaria brassicae* in seeds of rapeseed and mustard, its location in seeds, transmission from seeds to seedlings and control. *Seed Science and Technology* 28, 75–84.

Siemens, J. (2002) Interspecific hybridisation between wild relatives and *Brassica napus* to introduce new resistance traits into the oilseed rape gene pool. *Czech Journal of Genetics and Plant Breeding* 38, 155–157.

Siemens, J. and Sacristan, M.D. (1994) Asymmetric fusion between *Arabidopsis thaliana* and *Brassica nigra*. *ISHS Symposium on Brassicas, 9th Crucifer Genetics Workshop*, Lisbon, Portugal, 15–18 November 1994, pp. 181–192.

Siemens, J. and Sacristan, M.D. (1995) Production and characterization of somatic hybrids between *Arabidopsis thaliana* and *Brassica nigra*. *Plant Science* 111, 95–106.

Sigareva, M. and Earle, E.D. (1997a) Intertribal somatic hybrids between *Camelina sativa* and rapid cycling *Brassica oleracea*. *Cruciferae Newsletter* 19, 49–50.

Sigareva, M. and Earle, E.D. (1997b). *Capsella bursa-pastoris*: Regeneration of plants from protoplasts and somatic hybridization with rapid cycling *Brassica oleracea*. *Cruciferae Newsletter* 19, 57–58.

Sigareva, M.A. and Earle, E.D. (1999a) Camalexin induction in intertribal somatic hybrids between *Camelina sativa* and rapid-cycling *Brassica oleracea*. *Theoretical and Applied Genetics* 98, 164–170.

Sigareva, M.A. and Earle, E.D. (1999b) Regeneration of plants from protoplasts of *Capsella bursa-pastoris* and somatic hybridization with rapid cycling *Brassica oleracea*. *Plant Cell Reports* 18, 412–417.

Sikdar, S.R., Chatterjee, G., Das, S. and Sen, S.K. (1990) 'Erussica', the intergeneric fertile somatic hybrid developed through protoplast fusion between *Eruca sativa* Lam. and *Brassica juncea* (L.) Czern. *Theoretical and Applied Genetics* 79, 561–567.

Singh, M.P. and Kolte, S.J. (1999) Differential reactions of various crucifer host species against isolates of *Peronospora parasitica*. *Journal of Mycology and Plant Pathology* 29, 118–121.

Singh, R., Ellis, P.R., Pink, D.A.C. and Phelps, K. (1994) An investigation of the resistance to cabbage aphid in *Brassica* species. *Annals of Applied Biology* 125, 457–465.

Singh, S., Kaur, G., Gupta, M., Banga, S. and Banga, S.S. (2016) Genomic affinities between *Brassica napus* and *Raphanus raphanistrum* as revealed by meiotic GISH. *Plant Breeding* 135, 459–465.

Singh, S.P. and Sachan, G.C. (1997) Effect of different temperatures and host plants on the developmental behaviour of mustard sawfly, *Athalia proxima*. *Indian Journal of Entomology* 59, 34–40.

Sjodin, C. and Glimelius, K. (1989a) *Brassica naponigra*, a somatic hybrid resistant to *Phoma lingam*. *Theoretical and Applied Genetics* 77, 651–656.

Sjodin, C. and Glimelius, K. (1989b) Transfer of resistance against *Phoma lingam* to *Brassica napus* by asymmetric somatic hybridization combined with toxin selection. *Theoretical and Applied Genetics* 78, 513–520.

Skarzhinskaya, M., Landgren, M. and Glimelius, K. (1996) Production of intertribal somatic hybrids between *Brassica napus* L. and *Lesquerella fendleri* (Gray) Wats. *Theoretical and Applied Genetics* 93, 1242–1250.

Soroka, J., Gugel, R., Elliott, R., Rakow, G. and Raney, J.P. (2003) Resistance of crucifer species to insect pests. *Proceedings of the 11th GCIRC International Rapeseed Congress*, Copenhagen, Denmark, 6–10 July 2003, pp. 1031–1033.

Sosnowska, K. and Cegielska-Taras, T. (2014) Application of in vitro pollination of opened ovaries to obtain *Brassica oleracea* L. × *B. rapa* L. hybrids. *In Vitro Cellular & Developmental Biology – Plant* 50, 257–262.

Sridevi, O. and Sarla, N. (2005) Production of intergeneric hybrids between *Sinapis alba* and *Brassica carinata*. *Genetic Resources and Crop Evolution* 52, 839–845.

Srinivasan, K., Malathi, V.G., Kirti, P.B., Prakash, S. and Chopra, V.L. (1998) Generation and characterisation of monosomic chromosome addition lines of *Brassica campestris–B. oxyrrhina*. *Theoretical and Applied Genetics* 97, 976–981.

Stanek, R. and Lipecki, J. (1991) Resistance to triazines of *Capsella bursa-pastoris* (L.) Med. is located in chloroplast. *Resistant Pest Management* 3, 27.

Stebbins, G.L. (1958) The inviability, weakness, and sterility of interspecific hybrids. *Advances in Genetics* 9, 147–215.

Stebbins, G. L. (1966) Chromosomal variation and evolution. *Science* 152, 1463–1469.

Stiewe, G. and Robbelen, G. (1994) Establishing cytoplasmic male sterility in *Brassica napus* by mitochondrial recombination with *B. tournefortii*. *Plant Breeding* 113, 294–304.

Subrahmanyam, N.C. and Kasha, K.J. (1973) Selective chromosomal elimination during haploid formation in barley following interspecific hybridization. *Chromosoma* 42, 111–125.

Sundberg, E. and Glimelius, K. (1986) A method for production of interspecific hybrids within Brassiceae via somatic hybridization, using resynthesis of *Brassica napus* as a model. *Plant Science* 43, 155–162.

Sundberg, E. and Glimelius, K. (1991) Effects of parental ploidy level and genetic divergence on chromosome elimination and chloroplast segregation in somatic hybrids within Brassicaceae. *Theoretical and Applied Genetics* 83, 81–88.

Takashima, M., Bang, S. and Kaneko, Y. (2012) Production and characterization of monosomic addition lines of autoplasmic and alloplasmic *Brassica napus* with each B-genome chromosome of *Brassica juncea*. *Breeding Research* 14, 95–105.

Takeuchi, H. and Higashiyama, T. (2012) A species-specific cluster of defensin-like genes encodes diffusible pollen tube attractants in *Arabidopsis*. *PLoS Biology* 10, e1001449. DOI: 10.1371/journal.pbio.1001449.

Tan, C., Cui, C., Xiang, Y., Ge, X. and Li, Z. (2017) Development of *Brassica oleracea–nigra* monosomic alien addition lines: Genotypic, cytological and morphological analyses. *Theoretical and Applied Genetics* 130, 2491–2504.

Tan, M.K., and Medd, R.W. (2002) Characterisation of the acetolactate synthase (ALS) gene of *Raphanus raphanistrum* L. and the molecular assay of mutations associated with herbicide resistance. *Plant Science* 163, 195–205.

Tanksley, S.D. and Nelson, J.C. (1996) Advanced backcross QTL analysis: A method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical and Applied Genetics* 92, 191–203.

Taskin, K.M., Turgut, K. and Scott, R.J. (2004) Apomictic development in *Arabis gunnisoniana*. *Israel Journal of Plant Sciences* 52, 155–160.

Terasawa, Y. (1932) Konstante amphidiploide Brassico-raphanus Bastarde. *Proceedings of the Imperial Academy* 8, 312–314.

Tewari, J.P., Bansal, V.K., Tewari, I., Stringam, G.R. and Thiagarajah, M.R. (1996) Reactions of some wild and cultivated accessions of *Eruca* against *Leptosphaeria maculans*. *Cruciferae Newsletter* 18, 130–131.

The Global Risks Report 2019 14th edition (2019) World Economic Forum, Geneva, Switzerland. Available at: http://www3.weforum.org/docs/WEF_Global_Risks_Report_2019.pdf (accessed 13 March 2019).

Thomas, H.M., Morgan, W.G., Meredith, M.R., Humphreys, M.W. and Leggett, J.M. (1994) Identification of parental and recombined chromosomes in hybrid derivatives of *Lolium multiflorum* × *Festuca pratensis* by genomic *in situ* hybridization. *Theoretical and Applied Genetics* 88, 909–913.

Thompson, K.F. (1963) Resistance to the cabbage aphid (*Brevicoryne brassicae*) in *Brassica* plants. *Nature* 198, 209.

Tingdong, F., Guangsheng, Y. and Xiaoniu, Y. (1990) Studies on 'Three Line' polima cytoplasmic male sterility developed in *Brassica napus* L. *Plant Breeding* 104, 115–120.

Tonosaki, K., Michiba, K., Bang, S.W., Kitashiba, H., Kaneko, Y., *et al*. (2013) Genetic analysis of hybrid seed formation ability of *Brassica rapa* in intergeneric crossings with *Raphanus sativus*. *Theoretical and Applied Genetics* 126, 837–846.

Toriyama, K., Hinata, K. and Kameya, T. (1987a) Production of somatic hybrid plants¦Brassicomoricandia¦, through protoplast fusion between *Moricandia arvensis* and *Brassica oleracea*. *Plant Science* 48, 123–128.

Toriyama, K., Kameya, T. and Hinata, K. (1987b). Selection of a universal hybridizer in *Sinapis turgida* Del. and regeneration of plantlets from somatic hybrids with *Brassica* species. *Planta* 170, 308–313.

Tovar-Mendez, A., Kumar, A., Kondo, K., Ashford, A., Baek, Y.S., *et al*. (2014) Restoring pistil-side self-incompatibility factors recapitulates an interspecific reproductive barrier between tomato species. *The Plant Journal* 77, 727–736.

Tsunoda, S. (1980) Ecophysiology of wild and cultivated forms in *Brassica* and allied genera. In: Tsunoda, S., Hinata, K. and Gomez-Campo, C. (eds) *Brassica Crops and Wild Allies: Biology and Breeding*. Japan Scientific Societies Press, Tokyo, pp. 102–109.

Tsutsui, K., Jeong, B.H., Ito, Y., Bang, S.W. and Kaneko, Y. (2011) Production and characterization of an alloplasmic and monosomic addition line of *Brassica rapa* carrying the cytoplasm and one chromosome of *Moricandia arvensis*. *Breeding Science* 61, 373–379.

Tu, Y., Sun, J., Liu, Y., Ge, X., Zhao, Z., *et al*. (2008) Production and characterization of intertribal somatic hybrids of *Raphanus sativus* and *Brassica rapa* with dye and medicinal plant *Isatis indigotica*. *Plant Cell Reports* 27, 873–883.

Tu, Y., Sun, J., Ge, X. and Li, Z. (2009) Chromosome elimination, addition and introgression in intertribal partial hybrids between *Brassica rapa* and *Isatis indigotica*. *Annals of Botany* 103, 1039–1048.

Udagawa, H., Ishimaru, Y., Li, F., Sato, Y., Kitashiba, H., *et al*. (2010) Genetic analysis of interspecific incompatibility in *Brassica rapa*. *Theoretical and Applied Genetics* 121, 689–696.

Ueno, O., Bang, S.W., Wada, Y., Kondo, A., Ishihara, K., *et al*. (2003) Structural and biochemical dissection of photorespiration in hybrids differing in genome constitution between *Diplotaxis tenuifolia* (C3-C4) and radish (C3). *Plant Physiology* 132, 1550–1559.

Ulmer, B.J. and Dosdall, L.M. (2006) Glucosinolate profile and oviposition behavior in relation to the susceptibilities of Brassicaceae to the cabbage seedpod weevil. *Entomologia Experimentalis et Applicata* 121, 203–213.

Uprety, D.C., Prakash, S. and Abrol, Y.P. (1995) Variability for photosynthesis in *Brassica* and allied genera. *Indian Journal of Plant Physiology* 38, 207–213.

Van Tuyl, J.M., Van Dien, M.P., Van Creij, M.G.M., Van Kleinwee, T.C.M., Franken, J., *et al*. (1991) Application of *in vitro* pollination, ovary culture, ovule culture and embryo rescue for overcoming incongruity barriers in interspecific *Lilium* crosses. *Plant Science* 74, 115–126.

Velasco, L., Goffman, F.D. and Becker, H.C. (1998) Variability for the fatty acid composition of the seed oil in a germplasm collection of the genus *Brassica*. *Genetic Resources and Crop Evolution* 45, 371–382.

Vollmann, J., Grausgruber, H., Stift, G., Dryzhyruk, V. and Lelley, T. (2005) Genetic diversity in camelina germplasm as revealed by seed quality characteristics and RAPD polymorphism. *Plant Breeding* 124, 446–453.

Voss, A., Snowdon, R.J. and Luhs, W. (2000) Intergeneric transfer of nematode resistance from *Raphanus sativus* into the *Brassica napus* genome. *Acta Horticulturae* 539, 129–134. DOI: 10.17660/ActaHortic.2000.539.16.

Vyas, P., Prakash, S. and Shivanna, K.R. (1995) Production of wide hybrids and backcross progenies between *Diplotaxis erucoides* and crop brassicas. *Theoretical and Applied Genetics* 90, 549–553.

Walsh, M.J., Powles, S.B., Beard, B.R., Parkin, B.T. and Porter, S.A. (2004) Multiple-herbicide resistance across four modes of action in wild radish (*Raphanus raphanistrum*). *Weed Science* 52, 8–13.

Walsh, M.J., Owen, M.J. and Powles, S.B. (2007) Frequency and distribution of herbicide resistance in *Raphanus raphanistrum* populations randomly collected across the Western Australian wheatbelt. *Weed Research* 47, 542–550.

Walters, T.W., Mutschler, M.A. and Earle, E.D. (1992) Protoplast fusion-derived Ogura male sterile cauliflower with cold tolerance. *Plant Cell Reports* 10, 624–628.

Wang, G.X., Tang, Y., Yan, H., Sheng, X.G., Hao, W.W., *et al*. (2011a) Production and characterization of interspecific somatic hybrids between *Brassica oleracea* var. *botrytis* and *B. nigra* and their progenies for the selection of advanced pre-breeding materials. *Plant Cell Reports* 30, 1811–1821.

Wang, G.X., Yan, H., Zeng, X.Y., Sheng, X.G., Tang, Y., *et al*. (2011b) New alien addition lines resistance to black rot generated by somatic hybridization between cauliflower and black mustard. *Acta Horticulturae Sinica* 38, 1901–1910.

Wang, J., Gao, Y.N., Kong, Y.Q., Jiang, J.J., Li, A.M., *et al*. (2014) Abortive process of a novel rapeseed cytoplasmic male sterility line derived from somatic hybrids between *Brassica napus* and *Sinapis alba*. *Journal of Integrative Agriculture* 13, 741–748.

Wang, Y.P., Lan, L.F., Li, X.F. and Luo, P. (1999) A preliminary assessment of some wild cruciferous oil plant in the western Sichuan of China and their utilization. *Cruciferae Newsletter* 21, 29–30.

Wang, Y.P., Tang, J.S., Chu, C.Q. and Tian, J. (2000) A preliminary study on the introduction and cultivation of *Crambe abyssinica* in China, an oil plant for industrial uses. *Industrial Crops and Products* 12, 47–52.

Wang, Y.P., Sonntag, K. and Rudloff, E. (2003) Development of rapeseed with high erucic acid content by asymmetric somatic hybridization between *Brassica napus* and *Crambe abyssinica*. *Theoretical and Applied Genetics* 106, 1147–1155.

Wang, Y.P., Snowdon, R.J., Rudloff, E., Wehling, P., Friedt, W., *et al*. (2004) Cytogenetic characterization and fae1 gene variation in progenies from asymmetric somatic hybrids between *Brassica napus* and *Crambe abyssinica*. *Genome* 47, 724–731.

Wang, Y.P., Sonntag, K., Rudloff, E. and Chen, J.M. (2005a) Intergeneric somatic hybridization between *Brassica napus* L. and *Sinapis alba* L. *Journal of Integrative Plant Biology* 47, 84–91.

Wang, Y.P., Zhao, X.X., Sonntag, K., Wehling, P. and Snowdon, R.J. (2005b) Behaviour of *Sinapis alba* chromosomes in a *Brassica napus* background revealed by genomic *in-situ* hybridization. *Chromosome Research* 13, 819–826.

Wang, Y., Sonntag, K., Rudloff, E., Wehling, P. and Snowdon, R.J. (2006) GISH analysis of disomic *Brassica napus – Crambe abyssinica* chromosome addition lines produced by microspore culture from monosomic addition lines. *Plant Cell Reports* 25, 35–40.

Warwick, S.I. and Black, L.D. (1991) Molecular systematics of *Brassica* and allied genera (subtribe *Brassicinae, Brassiceae*) – chloroplast genome and cytodeme congruence. *Theoretical and Applied Genetics* 82, 81–92.

Warwick, S.I., and Sauder, C.A. (2005) Phylogeny of tribe *Brassiceae* (*Brassicaceae*) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast *trnL* intron sequences. *Canadian Journal of Botany* 83, 467–483.

Warwick, S.I., Black, L.D. and Aguinagalde, I. (1992) Molecular systematics of *Brassica* and allied genera (Subtribe *Brassicinae, Brassiceae*) – chloroplast DNA variation in the genus *Diplotaxis*. *Theoretical and Applied Genetics* 83, 839–850.

Warwick, S.I., Beckie, H.J., Thomas, A.G. and McDonald, T. (2000) The biology of Canadian weeds. 8. *Sinapis arvensis*. L. (updated). *Canadian Journal of Plant Science* 80, 939–961.

Warwick, S.I., Gugel, R.K., Gomez-Campo, C. and James, T. (2007) Genetic variation in *Eruca vesicaria* (L.) Cav. *Plant Genetic Resources* 5, 142–153.

Warwick, S.I., Francis, A. and Gugel, R.K. (2009) *Guide to Wild Germplasm of Brassica and Allied Crops (tribe Brassiceae, Brassicaceae)*. Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada.

Watanabe, M., Takayama, S., Isogai, A. and Hinata, K. (2003) Recent progresses on self-incompatibility research in *Brassica* species. *Breeding Science* 53, 199–208.

Wei, W., Zhang, S., Li, J., Wang, L., Chen, B., *et al*. (2006) Analysis of F_1 hybrid and BC1 monosomic alien addition line plants from *Brassica oleracea* × *Sinapis alba* by GISH. *Chinese Science Bulletin* 51, 2872. DOI: 10.1007/s11434-006-2207-9.

Westman, A.L. and Dickson, M.H. (1998) Disease reaction to *Alternaria brassicicola* and *Xanthomonas campestris* pv. *campestris* in *Brassica nigra* and other weedy crucifers. *Cruciferae Newsletter* 20, 87–88.

Westman, A.L., Kresovich, S. and Dickson, M.H. (1999) Regional variation in *Brassica nigra* and other weedy crucifers for disease reaction to *Alternaria brassicicola* and *Xanthomonas campestris* pv. *campestris*. *Euphytica* 106, 253–259.

Williams, P.H. and Pound, G.S. (1963) Nature and inheritance of resistance to *Albugo candida* in radish. *Phytopathology* 53, 1150–1154.

Winter, H., Gaertig, S., Diestel, A. and Sacristan, M.D. (1999) Blackleg resistance of different origin transferred into *Brassica napus*. *Proceedings of the 10th GCIRC International Rapeseed Congress*, Canberra, Australia, 26–29 September 1999. Available at: http://www.regional.org.au/au/gcirc (accessed 12 February 2019).

Winter, H., Diestel, A., Gärtig, S., Krone, N., Sterenberg, K., *et al*. (2003) Transfer of new blackleg resistances into oilseed rape. *Proceedings of the 11th GCIRC International Rapeseed Congress*, Copenhagen, Denmark, 6–10 July 2003, pp. 19–21.

Wu, J.G., Li, Z., Liu, Y., Liu, H.L. and Fu, T.D. (1997) Cytogenetics and morphology of the pentaploid hybrid between *Brassica napus* and *Orychophragmus violaceus* and its progeny. *Plant Breeding* 116, 251–257.

Wu, Y., Li, P., Zhao, Y., Wang, J. and Wu, X. (2007) Study on photosynthetic characteristics of *Orychophragmus violaceus* related to shade-tolerance. *Scientia Horticulturae* 113, 173–176.

Wu, Y.V. and Hojilla-Evangelista, M.P. (2005) *Lesquerella fendleri* protein fractionation and characterization. *Journal of the American Oil Chemists' Society* 82, 53–56.

Xu, C.Y., Wan-Yan, R.H. and Li, Z.Y. (2007a) Origin of new *Brassica* types from a single intergeneric hybrid between *B. rapa* and *Orychophragmus violaceus* by rapid chromosome evolution and introgression. *Journal of Genetics* 86, 249–257.

Xu, C.Y., Zeng, X.Y. and Li, Z.Y. (2007b) Establishment and characterization of *Brassica juncea* – *Orychophragmus violaceus* additions, substitutions and introgressions. *Euphytica* 156, 203–211.

Yajima, W., Hall, J.C. and Kav, N.N. (2004) Proteome-level differences between auxinic-herbicide-susceptible and-resistant wild mustard (*Sinapis arvensis* L.). *Journal of Agricultural and Food Chemistry* 52, 5063–5070.

Yamagishi, H., Landgren, M., Forsberg, J. and Glimelius, K. (2002) Production of asymmetric hybrids between *Arabidopsis thaliana* and *Brassica napus* utilizing an efficient protoplast culture system. *Theoretical and Applied Genetics* 104, 959–964.

Yan, Z., Tian, Z., Huang, R., Huang, B. and Meng, J. (1999) Production of somatic hybrids between *Brassica oleracea* and the C3–C4 intermediate species *Moricandia nitens*. *Theoretical and Applied Genetics* 99, 1281–1286.

Yaniv, Z., Elber, Y., Zur, M. and Schafferman, D. (1991) Differences in fatty acid composition of oils of wild Cruciferae seed. *Phytochemistry* 30, 841–843.

Yaniv, Z., Schafferman, D., Elber, Y., Ben-Moshe, E. and Zur, M. (1994) Evaluation of *Sinapis alba*, native to Israel, as a rich source of erucic acid in seed oil. *Industrial Crops and Products* 2, 137–142.

Yaniv, Z., Elber, Y., Schafferman, D., Ben-Moshe, E. and Zur, M. (1995) A survey of crucifers native to Israel, as a source of oils. *Plant Genetic Resources Newsletter* 101, 1–5.

Yaniv, Z., Schafferman, D. and Amar, Z. (1998) Tradition, uses and biodiversity of rocket (*Eruca sativa*, *Brassicaceae*) in Israel. *Economic Botany* 52, 394–400.

Zanetti, F., Monti, A. and Berti, M.T. (2013) Challenges and opportunities for new industrial oilseed crops in EU-27: A review. *Industrial Crops and Products* 50, 580–595.

Zenkteler, M. (1990) In vitro fertilization of ovules of some species of Brassicaceae. *Plant Breeding* 105, 221–228.

Zenkteler, M. (1991) Ovule culture and test tube fertilization. *Med Fac Landbouww Rijksuniv Genet* 56, 1403–1410.

Zenkteler, M. (2000) In vitro pollination of angiosperm ovules with gymnosperm pollen grains. *In Vitro Cellular & Developmental Biology – Plant* 36, 125–127.

Zhang, X., Ge, X., Shao, Y., Sun, G. and Li, Z. (2013) Genomic change, retrotransposon mobilization and extensive cytosine methylation alteration in *Brassica napus* introgressions from two intertribal hybridizations. *PLoS ONE* 8, e56346. DOI: 10.1371/journal.pone.0056346.

Zhao, J., Udall, J.A., Quijada, P.A., Grau, C.R., Meng, J., *et al*. (2006) Quantitative trait loci for resistance to *Sclerotinia sclerotiorum* and its association with a homeologous non-reciprocal transposition in *Brassica napus* L. *Theoretical and Applied Genetics* 112, 509–516.

Zhao, Z., Ma, N. and Li, Z. (2007) Alteration of chromosome behavior and synchronization of parental chromosomes after successive generations in *Brassica napus* × *Orychophragmus violaceus* hybrids. *Genome* 50, 226–233.

Zhao, Z.G., Hu, T.T., Ge, X.H., Du, X.Z., Ding, L., *et al*. (2008) Production and characterization of intergeneric somatic hybrids between *Brassica napus* and *Orychophragmus violaceus* and their backcrossing progenies. *Plant Cell Reports* 27, 1611–1621.

Zubr, J. and Matthaus, B. (2002) Effects of growth conditions on fatty acids and tocopherols in *Camelina sativa* oil. *Industrial Crops and Products* 15, 155–162.

20 Biofortified Pearl Millet Cultivars Offer Potential Solution to Tackle Malnutrition in India

Mahalingam Govindaraj[1], Parminder S. Virk[2]*, Anand Kanatti[1], Binu Cherian[2], K.N. Rai[1], Meike S. Anderson[2] and Wolfgang H. Pfeiffer[2]

[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India; [2]HarvestPlus, c/o International Food Policy Research Institute (IFPRI), Washington DC, USA

Introduction

Dietary deficiency of micronutrients (iron, zinc, vitamin A), leading to micronutrient malnutrition or hidden hunger, has been recognized as a widespread food-related health problem, affecting more than 2 billion people worldwide (White and Broadley, 2009; Stein 2010; FAO, 2015; Darnton-Hill and Mkparu, 2015). This is primarily attributable to lack of affordability and access to diversified diet, such as fruits, vegetables and livestock products. As a consequence, women, children and infants, belonging to the poorer section of society are malnourished. In particular, deficiencies of iron and zinc are widespread, leading to numerous adverse health consequences, as they play a vital role in various physiological body functions.

India is a country historically plagued by malnutrition, where nearly 30% of people lie below the poverty line, with little dietary diversity, because of poverty and low purchasing power. Although government-supported programmes have shown a reduction in malnutrition across the decades, there has been slow progress, as National Family Health Survey-3 and National Family Health Survey-4 reported by IIPS and MI (2007) and IIPS and ICF (2017), respectively, revealed. In addition, there is an unacceptably high prevalence of anaemia, and underweight and stunted children under 5 years of age (Table 20.1). More than 50% of children and women in 20 states of India are reported to be anaemic (Fig. 20.1). This situation is further compounded because present diets are dominated by major fine cereals, such as rice and wheat, which are often low in micronutrients and are readily available in ready-to-cook forms through the Public Distribution System (PDS) at a subsidized price. The costs of these micronutrient deficiencies in preventable lives lost, poor quality of life and adverse health issues, as well as their impact on personal and national economic growth are huge, even in a country like India, which has commendable economic growth. Micronutrient deficiencies alone may cost India US$2.5 billion annually (Gragnolati *et al*., 2005) and productivity loss of almost 3% GDP (Horton, 1999). Malnutrition, therefore, remains a serious problem in India, which is not only a consequence of poverty but also a cause of poverty (IFPRI, 2011).

Therefore, a multidisciplinary, sustainable and cost-effective approach, dubbed 'Biofortification –

* Email: p.virk@cgiar.org

© CAB International, 2020. *Quantitative Genetics, Genomics and Plant Breeding*, 2nd Edition (ed. M.S. Kang)

Table 20.1. Key indicators (in per cent) and magnitude of malnutrition in India over the years (IIPS and Macro International, 2007; IIPS and ICF, 2017).

Vulnerable group	NFHS-3 (2005–06)	NFHS-4 (2015–16)
Anaemia[a]		
Children (<5 years)	69.4	58.4
Non-pregnant women	55.2	53.1
Pregnant women	57.9	50.3
All women	55.3	53.0
All Men	24.2	22.7
Stunting		
Children	48.0	38.4
Women	NA[b]	NA
Men	NA	NA
Underweight		
Children	42.5	35.7
Women	35.5	22.9
Men	34.2	20.2

[a]Haemoglobin in grams per decilitre (g dl^{-1}).
[b]NA – not applicable.

breeding of staple crops for micronutrients', is ongoing both at global and national levels to bring together the potential of crop breeding and nutrition science to address hidden hunger. This effort is led by HarvestPlus – a CGIAR Challenge Programme convened by the International Food Policy Research Institute (IFPRI). In India, to start with, we chose to biofortify pearl millet (*Pennisetum glaucum* (L.) R. Br.), which is already known to be highly nutritious.

Why Biofortify Pearl Millet?

Pearl millet is grown in marginal arid and semi-arid tropical regions. India has the largest area in the world (8–9 million ha) and production (9–10 metric tonnes) (AICPMIP, 2016). It is cultivated on the sandiest, infertile soils, where other cereal crops fail to produce optimum yield. Being a dryland resilient crop and with high metabolizable energy, high protein content and balanced amino acid profile (Andrews and Kumar, 1992), pearl millet has a lot to offer. Low glycaemic index and gluten-free protein add special health benefits to pearl millet for those prone to diabetes and coeliac disease (Sehgal *et al*., 2004; Dahlberg *et al*., 2004). The consumption of pearl millet is higher in rural India, as compared to the urban population. This may be due to the fact that whole pearl millet grains are stored in villages, for up to 2 years, with few shelf-life challenges. However the shelf-life of pearl millet flour, as is stored by urban consumers, is short (10–12 days) because of rapid development of rancidity at ambient conditions (Satyavathi *et al*., 2017). Nevertheless, pearl millet continues to be an important staple for the poor and low-income groups. Pearl millet as such is a high-iron crop with a fairly high zinc content. However, not all available cultivars have high iron and zinc content. For instance, nearly twofold variability (31–61 mg kg^{-1}) was observed for iron content and one-and-a-half-fold variability (32–54 mg kg^{-1}) for Zn content among 122 commercial and pipeline hybrids developed so far in India. The average level of iron in these commercial cultivars is 42 mg kg^{-1}; thus, there is a need to increase the Fe and Zn levels in this crop (Rai *et al*., 2016). Remarkably, pearl millet has larger variability for iron and zinc content than do rice and wheat. For instance, pearl millet has 300% and 600% higher iron content than in wheat and rice, respectively (Passi and Jain, 2014). This indicates that pearl millet is a suitable target crop for iron biofortification. Unlike other crops, pearl millet foods are prepared using wholegrain flour and no significant losses occur in the total nutrient content during processing. Considering these adaptive and nutritional features, combined with high yield potential, pearl millet is an important cereal crop that can effectively address the emerging challenges of climate change, water scarcity for agriculture and food-related health issues, particularly iron-deficiency-induced anaemia. In addition, HarvestPlus developed the Biofortification Priority Index (BPI) to help stakeholders assess which country–crop combinations will have the greatest impact in reducing micronutrient deficiencies. The BPI ranks 128 countries according to their impact potential for investment in each of the eight biofortified staple food crops (Asare-Marfo *et al*., 2013; https://bpi.harvestplus.org). India ranks first for intervention with iron-biofortified pearl millet under the population-weighted BPI and second under the area-weighted BPI (Fig. 20.2). Therefore, expanding pearl millet's role as a biofortified food in dryland systems is highly important in the research and development sectors.

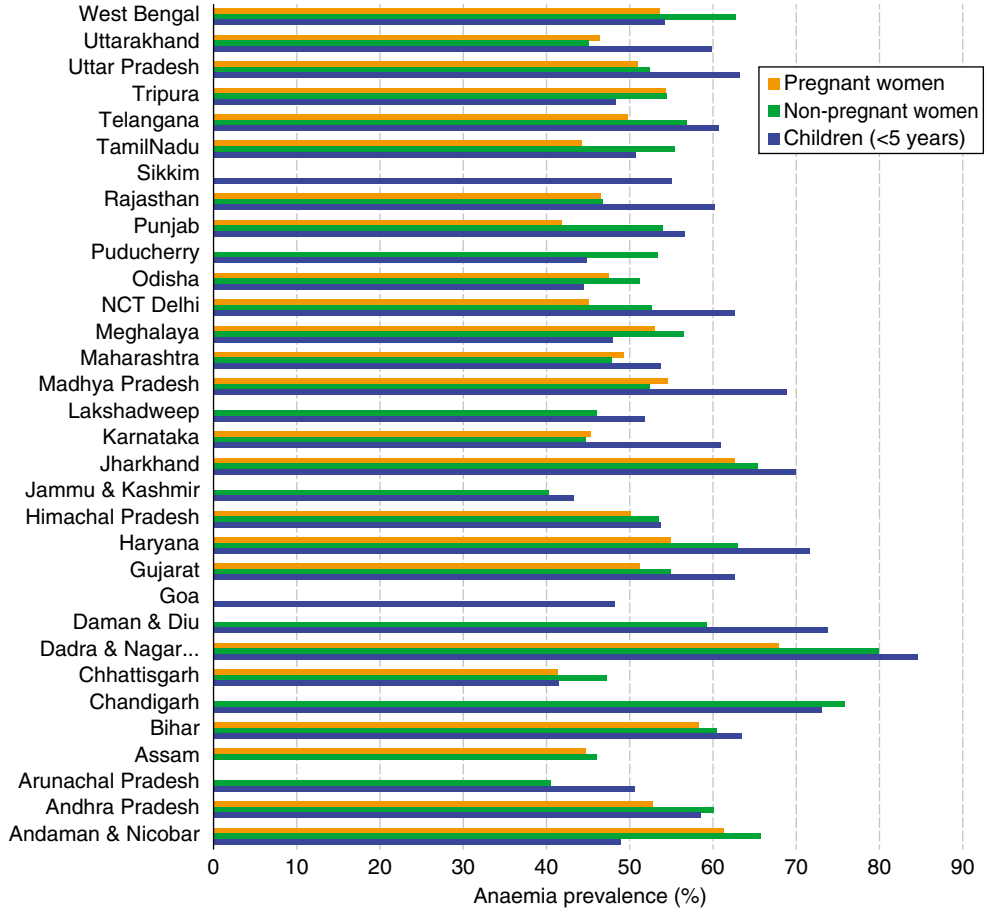


Fig. 20.1. High prevalence of anaemia (>40%) among vulnerable groups in different states across India. (IIPS and ICF, 2017.)

Breeding Target for Biofortified Pearl Millet

The primary target trait in pearl millet breeding is iron content, while zinc is an associated secondary trait. Nutritionists at HarvestPlus agreed that to detect measurable health impact, >30% of the estimated average requirement (EAR) should be achieved through biofortification. The baseline for iron for wholegrain pearl millet was found to be 47 mg kg^{-1} and an additional 30 mg kg^{-1} over the baseline was set as a breeding target (i.e. 77 mg kg^{-1}). These figures were arrived at on the basis of per capita consumption (220 g day^{-1} for adult women; 85 g day^{-1} for children 4–6 years of age), 7.5% bioavailability and 90% micronutrient retention after processing (Pfeiffer *et al.*, 2018).

On the other hand, the baseline for Fe was revisited using more extensive commercial hybrid trial data in India, to derive a baseline for hybrids that occupy more than 90% of the pearl millet area under improved cultivars in India. In an extensive study, Rai *et al.* (2016) suggested a baseline of 42 mg kg^{-1} in hybrids, hence the revised target level of iron for hybrids is set at 72 mg kg^{-1}. It is interesting to note that all the high-Fe cultivars (both open-pollinated varieties [OPVs] and hybrids) identified in this study have ≥35 mg kg^{-1} of Zn density, which is on a par with the target level determined for biofortified high-Zn wheat varieties (Velu *et al.*, 2012).

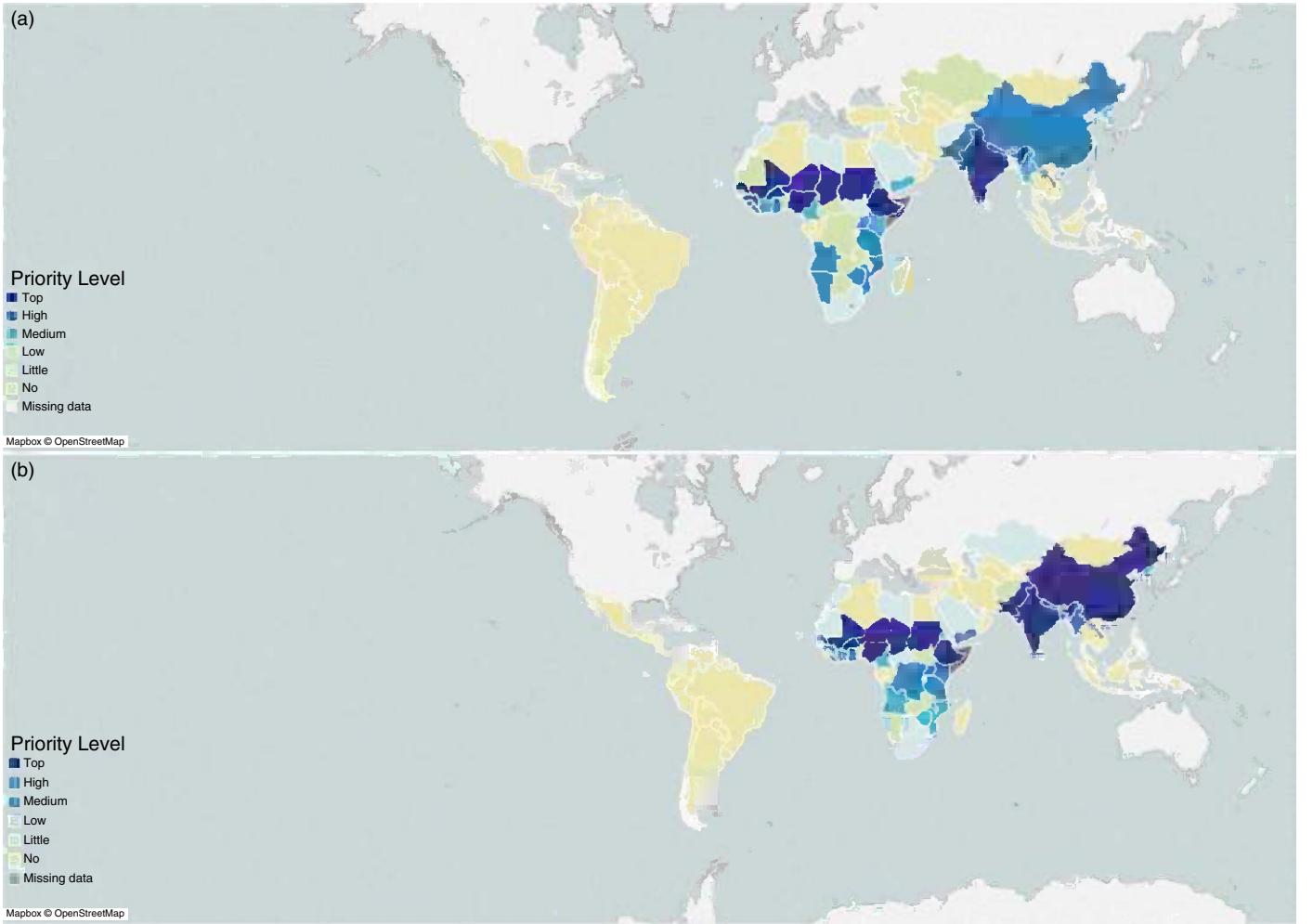


Fig. 20.2. Biofortification priority index indicating significance of breeding iron pearl millet. (a) Priority countries based on population weighted; (b) priority countries based on land area weighted.

Improved Phenotyping Protocol

For the success of a breeding programme, high-throughput, cost-effective phenotyping is a prerequisite. Initially, phenotyping was possible with the support from a well-established analytical laboratory facility at ICRISAT, which does high precision analysis following atomic absorption spectrometer (AAS) and inductively coupled plasma (ICP) procedures. However, these methods are very expensive and offer very low throughput. Hence, there was a need to develop a cost-effective and high-throughput screening procedure to speed up the breeding and product development process. The X-ray fluorescence (XRF) spectrometry calibrations and standards were developed for high-throughput screening in pearl millet (Paltridge *et al.*, 2012). The XRF method is low-cost, non-destructive, needs a small quantity of sample and has the ability to screen >300 samples per day (Rai *et al.*, 2012; Govindaraj *et al.*, 2016a). As pearl millet is a cross-pollinated crop, three types of seed samples (selfed, sibbed and open-pollinated [OP]) can be used for mineral analysis. The differences for iron and zinc among these types of grain samples were not significant (Rai *et al.*, 2015a), suggesting that the utilization of OP seed is the most cost-effective method of estimating Fe and Zn density. Research on genetic enhancement of grain iron and zinc content in pearl millet at ICRISAT has made significant progress in assessing the variability for these micronutrients in germplasm accessions and breeding lines using the above protocols.

Cultivar Development Strategy

In India, pearl millet OPVs were the dominant cultivars in the past (still occupying ≈ 30% area). However, now hybrids are cultivated on more than 5 million ha. The biofortification breeding at ICRISAT has assumed full operational scale and breeding pipeline, including OPVs, hybrids and hybrid-parent development. Both the public and private sectors are actively engaged in breeding pearl millet hybrids. This initiative is supported by ICRISAT through the Pearl Millet Hybrid Parent Research Consortium (PMHPRC).

The extent of genetic variation is very critical to initiate a breeding programme aimed at trait-specific breeding. The assessment of micronutrient variation was undertaken using phenotyping protocols, as described earlier. Pearl millet showed large genetic variability for both Fe and Zn densities in advanced breeding lines, populations and germplasm (Govindaraj *et al.*, 2015; Table 20.2), indicating good prospects for their genetic enhancement. As these traits were governed by additive gene action and their heritabilities were relatively high (Velu *et al.*, 2011; Govindaraj *et al.*, 2013; Kanatti *et al.*, 2014; Govindaraj *et al.*, 2016b), the pedigree method of breeding was deployed for progenies derived from primarily biparental crosses, as described by Andrews *et al.* (1996). It also meant that hybrid parental lines should be bred separately for high micronutrient density, requiring a separate hybrid parent-development programme.

Table 20.2. Variability for iron (Fe) and zinc (Zn) densities in pearl millet at ICRISAT, Patancheru. (Govindaraj *et al.*, 2015.)

Material	Entry	Per cent entry in micronutrient class (mg kg^{-1})							
		Fe density							
		≤45	46–55	56–65	66–75	76–85	86–95	96–105	>105
Mainstream hybrid parents	290	24	31	18	10	7	7	3	1
Commercial cultivars	140	56	35	9	1	0	0	0	0
Germplasm accessions	406	11	19	16	19	17	12	5	1
Biofortified breeding lines	514	0	0	0	2	11	22	27	38
		Zn density							
		≤35	36–45	46–55	56–65	66–75	76–85	86–95	>95
Mainstream hybrid parents	290	5	47	34	11	2	0	0	0
Commercial cultivars	140	8	76	16	0	0	0	0	0
Germplasm accessions	406	2	16	31	32	17	2	0	0
Biofortified breeding lines	514	8	45	40	7	0	0	0	0

Almost all iron sources identified are based on *iniadi* germplasm (early-maturing, large-seeded landrace materials from a geographic area adjoining Togo, Ghana, Burkina Faso and Benin) or have a large proportion of *iniadi* germplasm in their parentage (Rai *et al*., 2015b). Hence, *iniadi* is a valuable germplasm resource for genetic improvement of micronutrients in pearl millet. Highly significant and positive correlations between iron and zinc content indicated good prospects for simultaneous selection for both micronutrients. Both micronutrients, in general, have been found not to be correlated with 1000-grain weight and flowering time, indicating that pearl millet cultivars with high Fe and Zn densities can be effectively bred with large grain size and in a range of maturity classes for different agroecological regions (Rai *et al*., 2014; Govindaraj *et al*., 2019). The major focus of the breeding programme is to develop higher yielding, high-iron hybrids with stable yield and enhanced iron, for the different agro-ecological zones in India. Major traits in delivering final biofortified products include resistance to diseases, such as downy mildew and blast, drought tolerance and fodder yield.

Current Status and Future Prospects for Biofortified Pearl Millet Cultivars

ICRISAT, in association with national partners, developed and identified a high-iron variety 'Dhanashakti' that had the highest level of iron content among all pearl millet cultivars produced so far. Dhanashakti was initially targeted for Maharashtra state, but it also performed equally well in other states of central and southern India and was released by Mahatma Phule Krishi Vidyapeeth for cultivation in all pearl millet-growing states of India in 2014 (Rai *et al*., 2014). It was developed by an intra-population improvement of ICTP 8203 that was released in 1990. Dhanashakti has 9% higher iron and 11% higher yield than ICTP 8203. ICMV 221, another popular OPV variety, is also under improvement for iron (Govindaraj *et al*., 2019).

Two hybrids (ICMH 1202 and ICMH 1203) were released in 2017 and notified by the All India Coordinated Millet Improvement Project (AICPMIP) in 2018. In 2018, five more hybrids, namely AHB 1269, HHB 311, RHB 233, RHB 234 and ICMH1301, were released in India (Table 20.3). Currently, five hybrids, namely GHB 1225, PBH 1625, AHB 1382, ICMH 1601 and RHB 257, are in various stages of evaluation in the AICPMIP or state trials.

The seeds of Dhanashakti had been produced and marketed by HarvestPlus partners, namely Nirmal Seeds, Maharashtra State Seed Corporation and Karnataka State Seed Corporation. The commercial production of ICMH 1201 was undertaken by Shaktivardhak Seed Company under the brand name 'Shakti-1201' and 13 metric tonnes of Truthfully Labelled Seed (TLS) was sold in the states of Maharashtra and Rajasthan (Purushottam Singh *et al*., 2016). Adoption of recently released hybrids is likely to take place in the near future. Overall, >93,000 households in four states (Maharashtra, Rajasthan, Uttar Pradesh and Haryana) have access to biofortified pearl

Table 20.3. Performance and some salient features of released biofortified pearl millet cultivars in India

Hybrid	Release year	Grain colour	Grain size	Yield potential (t ha^{-1})[a]	Iron density (mg kg^{-1})[b]
Dhanashakti	2014	Dark grey	Bold	2.0	71
ICMH 1202 (AHB1200Fe)	2017	Grey	Bold	3.5	70
ICMH 1203 (HHB 299)	2017	Grey	Bold	3.2	67
ICMH 1301 (DHBH1211)	2018	Grey	Bold	3.3	78
ICMH 1501 (HHB 311)	2018	Grey	Medium	3.5	60
ICMH 1502 (AHB1269)	2018	Grey	Bold	3.2	73
ICMH 1503 (RHB 233)	2018	Grey	Bold	3.2	65
ICMH 1504 (RHB 234)	2018	Grey	Medium	3.2	60

[a]Mean data from AICRP-PM test locations.
[b]XRF data.

millet since the first biofortified variety was released in India (Binu Cherian, HarvestPlus-ICRISAT, 2019; personal communication).

Seven hybrids released so far represent the first wave of hybrids that were developed in collaboration with national partners. These hybrids possess >70% of the iron target increments and about 10–20% lower grain yield than the highest yielding commercial hybrid check (Fig. 20.3, Table 20.3). This is primarily attributable to the fact that all the high-Fe pearl millet hybrids released so far were developed using pre-existing medium- to high-Fe seed parents and advanced high-Fe breeding lines as potential restorers. These materials were not purposely bred for high yield and Fe and Zn density as target traits in the mainstream breeding programme. Consequently, yield levels of the first wave of biofortified hybrids are not as high as those of commercial high-yielding checks (Govindaraj *et al.*, 2019).

However, the next wave of biofortified hybrids are being bred through directed breeding for high yield, along with high-Fe/Zn (Figure 20.4). In addition, to fulfill the long-term objective and continued supplies of breeding material, at ICRSIAT, we continue to mainstream breeding for iron and zinc, along with other core traits. The next generation of biofortified parental lines (with disease resistance) are being developed. The current breeding strategy ensures that the crosses include at least one parent having high iron content (>60 mg kg^{-1}) to enable mainstreaming of the iron trait in the next few years in a sustainable manner, as committed by the CGIAR Consortium and its member institutes at the Second Global Conference on Biofortification in Kigali, Rwanda. To this effect, promising lines have been identified with >80 mg kg^{-1} Fe density and 40–60 mg kg^{-1} Zn density and are being used in developing the next generation of seed-parents and restorer lines to develop the next generation of high-yielding and high-Fe hybrids. About 30 seed-parents (A/B pairs), purposely bred for high-Fe (70–110 mg kg^{-1}), have been designated with disease resistance in three diverse cytoplasm sources ($A_1/A_4/A_5$). Similarly, about 30 high-Fe restorers have also been identified and designated. Seeds of these designated parents are being multiplied for sharing with collaborators and will serve as potential parents of the future biofortified hybrids. We believe that this targeted

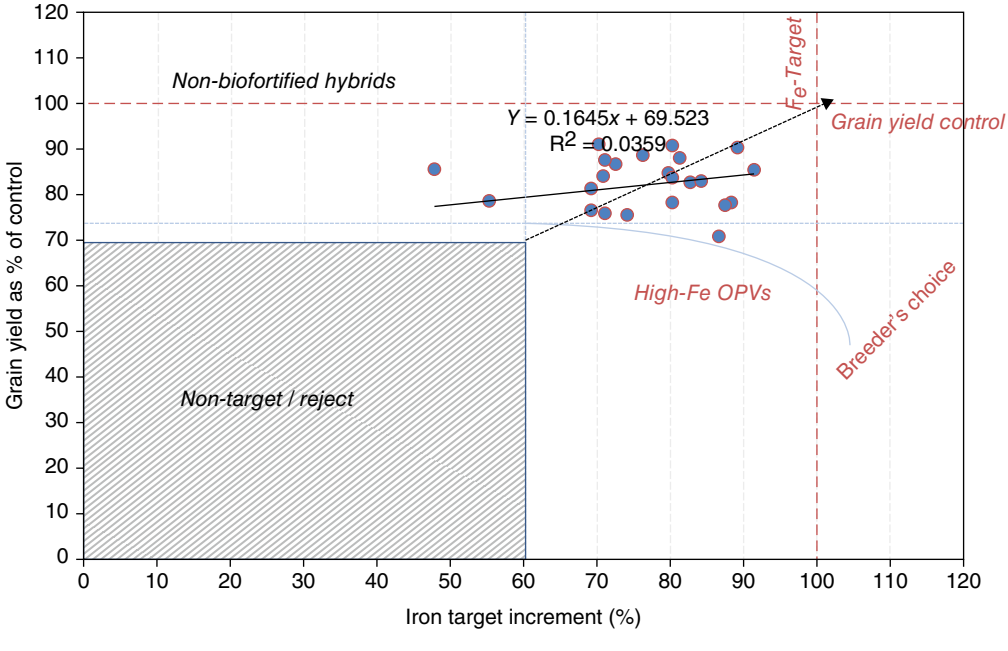


Fig. 20.3. Current biofortification breeding progress on hybrid development towards the target increments for yield and iron content (data from 2012 trials). OPV, open-pollinated variety.

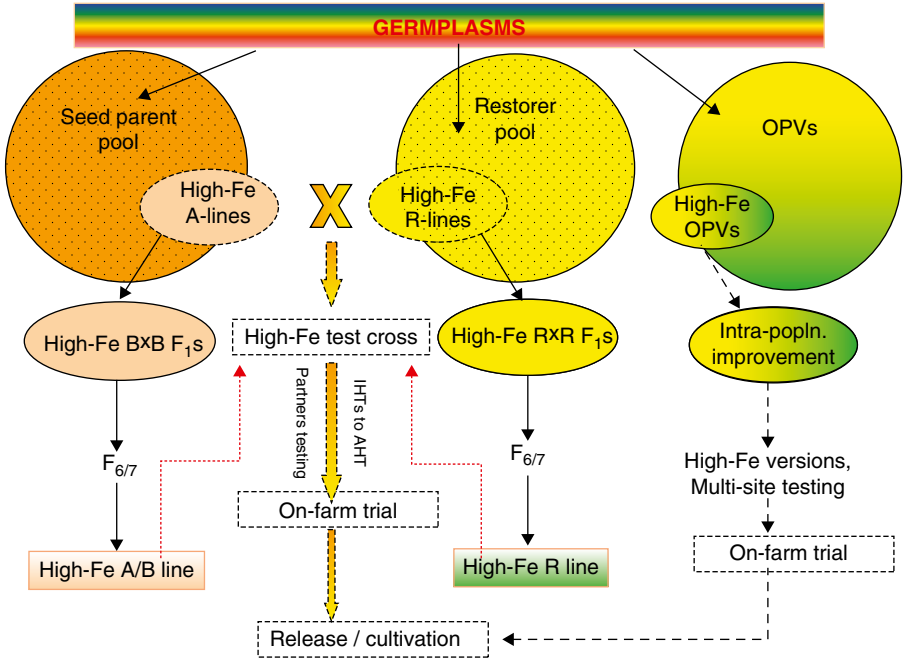


Fig. 20.4. Current fast-track biofortification breeding strategy (dotted lines) followed at ICRISAT. (Adapted from Govindaraj *et al*., 2019.) OPV, open-pollinated variety; IHTs, Initial hybrid trial(s); AHT, advanced hybrid trial.

breeding approach will result in the production of a second generation of biofortified hybrids with competitive yield potential comparable to or surpassing the commercial hybrids.

Biofortified Cultivar Release Policy

The Indian Council of Agricultural Research (ICAR) and AICPMIP have been very supportive of biofortification of pearl millet. As early as 2015, AICPMIP constructed a special module to test and release biofortified pearl millet cultivars in India. Furthermore, ICAR has endorsed the inclusion of the minimum levels of iron and zinc for future pearl millet varieties to be released in the country. AICPMIP in 2018 decided on a minimum of 42 mg kg^{-1} of iron and 32 mg kg^{-1} of zinc (AICRP-PM, 2017) (https://www.icrisat.org/committed-to-alleviating-malnutrition-india-declares-minimum-levels-for-iron-and-zinc-in-pearl-millet/). These mandatory initiatives should help mainstream biofortified traits in various national breeding programmes in India.

Nutritional Bioavailability and Efficacy Evidence

To demonstrate that the micronutrients bred into pearl millet or any other cereal are bioavailable and have measurable impact on micronutrient status and functional indicators of micronutrient status, such as physical activity and cognition tests, nutritionists conduct randomized, controlled efficacy trials. This evidence is necessary to convince all stakeholders, especially policy makers of the benefits of biofortified foods.

A set of efficacy studies was conducted (Kodkany *et al*., 2013; Pompano *et al*., 2013; Scott *et al*., 2014; Tako *et al*., 2015; Finkelstein *et al*., 2015, 2017, 2019; Jen Foley, 2019, Vancouver, personal communication). Research findings from these studies are briefly presented below.

- Pearl millet biofortified with iron can provide young children with 100% of their daily iron needs. When eaten as the main part of the diet, it reverses iron deficiency and improves the health of adolescent Indian children.

  - In rural Maharashtra, when secondary school children ate biofortified *bhakri* twice a day and snacks at will for 6 months, iron deficiency was significantly reduced.
  - Even after only 4 months, iron levels were significantly improved in all children eating iron pearl millet.
  - Notably, the impact of the additional dietary iron was greatest in those who needed it the most. Children with iron deficiency at the start of the study were 64% more likely to have their deficiency reversed after 6 months if they ate iron pearl millet instead of traditional millet.
- Eating iron pearl millet also improves cognitive performance, including memory and attention – skills essential for reaching one's full potential at school and work.
- Significant positive and protective effects from iron-biofortified crops, including iron pearl millet, have been proven across continents and populations, particularly among women and children in poor communities.
- In addition to increasing iron, biofortification increases zinc in pearl millet to levels that can meet the daily needs of young children. This is beneficial because iron, zinc and other micronutrient deficiencies often co-exist in the Indian population.

The iron-nutrition research presented above has demonstrated beyond doubt the efficacy of biofortified pearl millet in improving the nutritional status of target populations. Therefore, pearl millet biofortification offers a sustainable solution to iron and zinc deficiencies among millet-growing and millet-consuming populations and then can penetrate into non-millet growing but millet-consuming urban-poor populations for improving their health.

The Way Forward to Eradicate Malnutrition in India

Biofortification is one of the key approaches, which is cost-effective and sustainable to address global hidden hunger in the world, including India. The recent 'National Nutrition Strategy' by NITI Aayog, Government of India, would provide impetus to utilize biofortified varieties more effectively towards achieving 'Kuposhan Mukt Bharat' ('malnutrition-free India'). A framework has been suggested in the 'Vision 2022' for National Actions to eradicate malnutrition (Singh, 2019). The Integrated Child Development Service (ICDS) scheme – the world's largest nutrition programme, was launched in India in 1975 to address the health and nutrition needs of children under 6 years of age, which can be used as a vehicle to promote biofortified foods. Other initiatives, such as a Public Distribution System (PDS), Mid-Day Meal Scheme (MDMS) and the Food Bill, Food – a legal right, 2013, etc. are also in place to address micronutrient malnutrition. Innovative policy interventions, such as preferential seed subsidy and price incentives to grain producers and, would trigger adoption of iron pearl millet and other biofortified crops.

Summary and Conclusions

One-third of the global population suffers from one or more micronutrient deficiencies. More than 50% of children and women in 20 states of India are anaemic. Most of the suffering populations get their calories from main staple crops. Biofortification is the process of breeding micronutrient traits into staple food crops, which are bioavailable. It has been proven that biofortified crops have a positive and measurable impact on the health of consuming populations. Biofortification is a cost-effective and sustainable strategy and it complements the existing interventions, such as commercial food fortification and supplementation. However, biofortification has the potential to reach malnourished populations in relatively remote rural areas where other approaches have had little impact. Pearl millet, known to be highly nutritious, is the most important drought- and climate-resilient cereal crop, widely grown in Asia (9 million ha). It is a nutritious dryland crop having high protein, micronutrients and a more balanced amino acid profile than other staple cereals. In 2018, the Government of India, renamed millets, including pearl millet, as Nutri-Cereals. The average level of iron and zinc in the commercial pearl millet cultivars is 42 mg kg^{-1} and 32 mg kg^{-1}, respectively. However, the large extent of genetic variability for these traits (Fe 30–140 mg kg^{-1}, Zn 20–90 mg kg^{-1}) encouraged plant breeders to

improve their content in pearl millet. HarvestPlus and ICRISAT collaborated to address the Fe and Zn deficiencies through biofortification of pearl millet. However, the principal emphasis of pearl millet biofortification is on improving primarily grain Fe. A decade of research and development on high-iron breeding pipelines and release of eight biofortified pearl millet cultivars led to the cultivation of biofortified pearl millet varieties on ~75,000 ha in India, reaching >93,000 households. To scale up, investment commitment in crop breeding, both in the public and private sector, and favourable policies, such as seed subsidy, price incentives for grain producers and integration of biofortified varieties into the existing Public Distribution and Mid-Day Meal Schemes, are warranted.

Biofortification is an evidence-based, sustainable and cost-effective approach to address malnutrition through the development, release and adoption of yield-competitive varieties possessing additional micronutrient content (Bouis and Saltzman, 2017). Biofortification helps reach relatively remote rural populations, who have limited access to commercially marketed fortified foods. We recognize that biofortification and commercial fortification are complementary strategies. However, biofortification is particularly advantageous where households consume large amounts of food staples that are often poor in micronutrients, and are most vulnerable to hidden hunger (Bouis, 2000; Bouis *et al*., 2011). It has been shown that farmers have adopted both OPVs and hybrid varieties of pearl millet with higher iron content. However, we need to work with all stakeholders in the full value chain to maximize impact with a mindset of 'farm to fork' rather than just development and release of biofortified varieties. Hence, to scale up and strengthen the breeding pipeline and uptake of these varieties, public and private partners need to work together at national and international levels. Agriculture investments and favourable policies to promote biofortification will enhance the availability of nutritious food to farmers and communities. Linking biofortification to the ongoing government initiatives, such as inclusion of biofortified pearl millet in the Public Distribution System and Mid-Day Meal Schemes to address malnutrition, would trigger demand. Increased demand for biofortified food would create market opportunities for farmers, thereby boosting their income. Pearl millet has the potential to make significant contributions to food and nutritional security in semi-arid regions of India.

References

AICPMIP (2016) Annual Report, All India Coordinated Pearl Millet Improvement Project. AICPMIP, Indian Council of Agricultural Research, Mandor, Rajasthan, India.

AICRP-PM (2017) *Proceedings of the 52nd Annual Group Meeting of ICAR – All India Coordinated Research Project on Pearl Millet*, Ludhiana, India, 28–30 April 2017.

Andrews, D.J. and Kumar, K.A. (1992) Pearl millet for food, feed, and forage. *Advances in Agronomy* 48, 89–139.

Andrews, D.J., Hanna, W.W., Rajewski, J.F. and Collins, V.P. (1996) Advances in grain pearl millet: Utilization and production research. In: Janick, J. and Whipkey, A. (eds) *Issues in New Crops and New Uses*. ASHS Press, Alexandria, Virginia, pp. 170–177.

Asare-Marfo, D., Birol, E., Gonzalez, C., Moursi, M., Perez, S., *et al*. (2013) Prioritizing countries for biofortification interventions using country-level data. HarvestPlus Working Paper no. 11. International Food Policy Research Institute (IFPRI), Washington DC. Available at: http://www.ifpri.org/publication/biofortification-priority-index (accessed 5 November 2019).

Bouis, H.E. (2000) Improving human nutrition through agriculture. *Food and Nutrition Bulletin* 21, 549–565.

Bouis, H.E. and Saltzman, A. (2017) Improving nutrition through biofortification: A review of evidence from HarvestPlus, 2003 through 2016. *Global Food Security* 12, 49–58.

Bouis, H.E., Hotz, C., McClafferty, B., Meenakshi, J.V. and Pfeiffer, W.H. (2011) Biofortification: A new tool to reduce micronutrient malnutrition. *Food and Nutrition Bulletin* 32, 31S–40S.

Dahlberg J.A., J.P. Wilson, T. Snyder. (2004) Sorghum and pearl millet: Health foods and industrial products in developed countries. *Alternative Uses of Sorghum and Pearl Millet in Asia: Proceedings of*

the Expert Meeting, ICRISAT, CFC Technical Paper No. 34, Patancheru, Andhra Pradesh, India, 1–4 July 2003, pp. 42–59.

Darnton-Hill, I. and Mkparu U.C. (2015) Micronutrients in pregnancy in low- and middle-income countries. *Nutrients* 7, 1744–1768.

FAO (2015) *The State of Food Insecurity in the World 2015*. FAO, Rome.

Finkelstein, J.L., Mehta, S., Udipi, S.A., Ghugre, P.S., Luna, S.V., *et al*. (2015) A randomised trial of iron-biofortified pearl millet in school children in India. *The Journal of Nutrition* 145(7), 1576–1581. DOI: 10.3945/jn.114.208009.

Finkelstein J.L., Haas, J.D. and Mehta S. (2017) Iron-biofortified staple food crops for improving iron status: A review of the current evidence. *Current Opinion in Biotechnology* 44,138–145.

Finkelstein, J.L., Fothergill, A., Hackl, L., Haas, J. and Mehta, S. (2019) Iron biofortification interventions to improve iron status and functional outcomes. *Proceedings of the Nutrition Society* 78, 197–207. DOI: 10.1017/S0029665118002847.

Govindaraj, M., Rai, K.N., Shanmugasundaram, P., Dwivedi, S.L., Sahrawat, K.L., *et al*. (2013) Combining ability and heterosis for grain iron and zinc densities in pearl millet. *Crop Science* 53, 507–517.

Govindaraj, M., Rai, K.N., Kanatti, A., Velu, G. and Shivade, H. (2015) Breeding high-iron pearl millet cultivars: Present status and future prospects. *2nd International Conference on Global Food Security*, Ithaca, NY, 11–14 October 2015.

Govindaraj, M., Rai, K.N., Pfeiffer, W.H., Kanatti, A. and Shivade, H. (2016a) Energy-dispersive X-ray fluorescence spectrometry for cost-effective and rapid screening of pearl millet germplasm and breeding lines for grain iron and zinc density. *Communications Soil Science Plant Analysis* 47, 2126–2134.

Govindaraj, M., Rai, K.N. and Shanmugasundaram, P. (2016b) Intra-population genetic variance for grain iron and zinc contents and agronomic traits in pearl millet. *Crop Journal* 4, 48–54.

Govindaraj, M., Rai, K.N., Cherian, B., Pfeiffer, W.H., Kanatti, A., *et al*. (2019) Breeding biofortified pearl millet varieties and hybrids to enhance millet markets for human nutrition. *Agriculture* 9, 106. DOI: 10.3390/agriculture9050106.

Gragnolati, M., Shekar, M., Gupta, M.D., Bredenkamp, C. and Lee, Y.K. (2005) India's undernourished children: A call for reform and action. Health, Nutrition and Population Division, World Bank, Washington DC.

Horton, S. (1999) Opportunities for investments in nutrition in low-income Asia. In: Hunt, J. and Quibria, M.G. (eds) *Investing in Child Nutrition in Asia, Asian Development Nutrition and Development Series 1*. Asian Development Bank, Manila, pp. 246–273.

IFPRI (2011) Leveraging agriculture for improving nutrition and health. *IFPRI 2020, Vision International Conference*, New Delhi, India, 10–12 February 2011.

International Institute for Population Sciences (IIPS) and ICF (2017) *National Family Health Survey (NFHS-4), 2015–16: India*. IIPS, Mumbai, India.

International Institute for Population Sciences (IIPS) and Macro International (2007) *National Family Health Survey (NFHS-3), 2005–06: India: Volume I*. IIPS, Mumbai, India.

Kanatti, A., Rai, K.N., Radhika, K., Govindaraj, M., Sahrawat, K.L., *et al*. (2014) Grain iron and zinc density in pearl millet: Combining ability, heterosis and association with grain yield and grain size. *SpringerPlus* 3, 763.

Kodkany, B.S., Bellad, R.M., Mahantshetti, N.S., Westcott, J.E., Krebs, N.F., *et al*. (2013) Biofortification of pearl millet with iron and zinc in a randomized controlled trial increases absorption of these minerals above physiologic requirements in young children. *Journal of Nutrition* 143, 1489–1493.

Paltridge, N.G., Palmer, L.J., Milham, P.J., Guild, G.E. and Stangoulis, J.C.R. (2012) Energy-dispersive X-ray fluorescence analysis of zinc and iron concentration in rice and pearl millet grain. *Plant and Soil* 361, 251–60.

Passi, S.J, and Jain, A. (2014) Millets: The nutrient rich counterparts of wheat and rice. Press Information Bureau, Government of India. Available at: http://pib.nic.in/newsite/mbErel.aspx?relid=106818 (accessed 5 November 2019).

Pfeiffer, W., Anderson, M., Govindaraj, M. Virk, P., Cherian, B., *et al*. (2018) Biofortification in underutilized staple crops for nutrition in Asia and Africa. In: *Regional Expert Consultation on Underutilized Crops for Food and Nutritional Security in Asia and the Pacific – Thematic, Strategic Papers and Country Status Reports*. Asia-Pacific Association of Agricultural Research Institutions (APAARI), Bangkok, Thailand, pp. 70–81.

Pompano, L.M., Przybyszewski, E.M., Udipi, S.A., Ghugre, P. and Haas, J.D. (2013) VO2 max improves in Indian school children after a feeding trail with iron biofortified pearl millet. *The FASEB Journal* 27 (issue 1 supplement), 845.28.

Purushottam Singh S.K. and Uddeen, R. (2016) Nutri-Farms for mitigating malnutrition in India. In: Singh, U., Praharaj, C.S., Singh, S.S. and Singh, N.P. (eds) *Biofortification of Food Crops*. Springer, New Delhi, pp. 461–477.

Rai, K.N., Govindaraj, M. and Rao, A.S. (2012) Genetic enhancement of grain iron and zinc content in pearl millet. *Quality Assurance and Safety of Crops & Foods* 4, 119–125.

Rai, K.N., Gupta, S.K., Sharma, R., Govindaraj, M., Rao, A.S., *et al*. (2014) Pearl millet breeding lines developed at ICRISAT: A reservoir of variability and useful source of non-target traits. *SAT eJournal* 1, 1–13.

Rai, K.N., Govindaraj, M., Pfeiffer, W.H. and Rao, A.S. (2015a) Seed set and xenia effects on grain iron and zinc density in pearl millet. *Crop Science* 55, 821–827.

Rai, K.N., Velu, G., Govindaraj, M., Upadhyaya, H.D., Rao, A.S., *et al*. (2015b) Iniadi pearl millet germplasm as a valuable genetic resource for high grain iron and zinc densities. *Plant Genetic Resources* 13, 75–82.

Rai, K.N., Yadav, O.P., Govindaraj, M., Pfeiffer, W.H., Yadav, H.P., *et al*. (2016) Grain iron and zinc densities in released and commercial cultivars of pearl millet. *Indian Journal of Agriculture Sciences* 86, 291–296.

Satyavathi, C.T., Shelly, P., Saikat, M., Chugh, L.K. and Asha, K. (2017) *Enhancing Demand of Pearl millet as Super Grain – Current Status and Way Forward*. ICAR All India Coordinated Research Project on Pearl Millet, Jodhpur, India.

Scott, S., Murray-Kolb, L., Wenger, M., Udipi, S., Ghugre, P., *et al*. (2014) Iron-biofortified pearl millet improves attentional function in Indian adolescents, a 6-month randomized controlled trial. *The FASEB Journal* 28 (Supplement 1), 619.2.

Sehgal, S., Kawatra, A. and Singh, G. (2004) Recent advances in pearl millet and sorghum processing and food product development. *Alternative Uses of Sorghum and Pearl Millet in Asia, Proceedings of the Expert Meeting, ICRISAT*. Patancheru, Andhra Pradesh, India, 1–4 July 2003, pp. 60–92.

Singh, R.B. (2019) Nutrition-sensitive agriculture and food systems to build zero hunger new India. In: Dwivedi, B.S., Singh, V.K., Singh, S,K., *et al*. (eds) *Proceeding of XIV Agricultural Science Congress*. National Academy of Agricultural Sciences, New Delhi, pp. 32–66.

Stein, A.J. (2010) Global impacts of human malnutrition. *Plant and Soil* 335, 133–154.

Tako, E., Reed, S.M., Budiman, J., Hart, J.J. and Glahn, R.P. (2015) Higher iron pearl millet (*Pennisetum glaucum* L.) provides more absorbable iron that is limited by increased polyphenolic content. *Nutrition Journal* 14, 11.

Velu, G., Rai, K.N., Muralidharan, V., Longvah, T. and J. Crossa, J. (2011) Gene effects and heterosis for grain iron and zinc density in pearl millet (*Pennisetum glaucum* (L.) R. Br). *Euphytica* 180, 251–259.

Velu, G., Singh, R.P., Huerta-Espino, J., Peña-Bautista, R.J., Arun, B., *et al*. (2012) Performance of biofortified spring wheat genotypes in target environments for grain zinc and iron concentrations. *Field Crops Research* 137, 261–267.

White, P.J. and Broadley, M.R. (2009) Biofortification of crops with seven mineral elements often lacking in human diets-iron, zinc, copper, calcium, magnesium, selenium and iodine. *New Phytologist* 182, 49–84.

Index

Note: Page numbers in **bold** type refer to **figures**
Page numbers in *italic* type refer to *tables*

CABI – who we are and what we do

This book is published by **CABI**, an international not-for-profit organisation that improves people's lives worldwide by providing information and applying scientific expertise to solve problems in agriculture and the environment.

CABI is also a global publisher producing key scientific publications, including world renowned databases, as well as compendia, books, ebooks and full text electronic resources. We publish content in a wide range of subject areas including: agriculture and crop science / animal and veterinary sciences / ecology and conservation / environmental science / horticulture and plant sciences / human health, food science and nutrition / international development / leisure and tourism.

The profits from CABI's publishing activities enable us to work with farming communities around the world, supporting them as they battle with poor soil, invasive species and pests and diseases, to improve their livelihoods and help provide food for an ever growing population.

CABI is an international intergovernmental organisation, and we gratefully acknowledge the core financial support from our member countries (and lead agencies) including:

UKaid from the British people

Ministry of Agriculture People's Republic of China

Australian Government Australian Centre for International Agricultural Research

Agriculture and Agri-Food Canada

Ministry of Foreign Affairs of the Netherlands

Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra

Swiss Agency for Development and Cooperation SDC

Discover more

To read more about CABI's work, please visit: **www.cabi.org**

Browse our books at: **www.cabi.org/bookshop**,
or explore our online products at: **www.cabi.org/publishing-products**

Interested in writing for CABI? Find our author guidelines here:
www.cabi.org/publishing-products/information-for-authors/

Quantitative Genetics, Genomics and Plant Breeding

2nd Edition

Edited by Manjit S. Kang

Since the first edition of this book was published in 2002, the field of quantitative genetics, genomics and breeding has changed markedly. In response, only four chapters have been updated for this new edition, while the remaining 16 chapters are entirely new.

This book presents state-of-the-art, authoritative chapters on contemporary issues in the broad areas of quantitative genetics, genomics and plant breeding. Section 1 (Chapters 2 to 12) emphasizes the application of genomics, and genome and epigenome editing techniques, in plant breeding; bioinformatics; quantitative trait loci mapping; and the latest approaches of examining and exploiting genotype–environment interactions. Section 2 (Chapters 13 to 20) represents the intersection of breeding, genetics and genomics. This section describes the use of cutting-edge molecular breeding and quantitative genetics techniques applied to wheat, rice, maize, root and tuber crops and pearl millet. Overall, the book focuses on using genomic information to help evaluate traits that can combat biotic/abiotic stresses, genome-wide association mapping, high-throughput genotyping/phenotyping, biofortification, use of big data, orphan crops, and gene editing techniques. The examples featured are taken from across crop science research and cover a wide geographical base.

This book contains:

- chapters by expert authors from six continents;
- state-of-the-art information on topical areas relative to crop improvement;
- coverage of genome-editing techniques.

This book is recommended for researchers, and both graduate and post-graduate students, in genetics, crop breeding, genomics and quantitative genetics.