eISBN: 978-1-68108-385-8
ISBN: 978-1-68108-386-5

eISSN: 2405-7711
ISSN: 2405-7703

ANTHROPOLOGY
CURRENT AND FUTURE
DEVELOPMENTS Volume 2

GENOMICS IN BIOLOGICAL ANTHROPOLOGY

NEW **CHALLENGES**, NEW **OPPORTUNITIES**

Editors:
Manuela Lima
Amanda Ramos
Cristina Santos

Bentham e Books

Anthropology: Current and Future Developments

(Volume 2)

Genomics in Biological Anthropology: New Challenges, New Opportunities

Edited by

Manuela Lima

Department of Biology,
University of the Azores,
Ponta Delgada,
Portugal

Amanda Ramos & Cristina Santos

Universitat Autònoma de Barcelona,
Cerdanyola del Vallès,
Barcelona,
Spain

Anthropology: Current and Future Developments

Volume # 2

Genomics in Biological Anthropology: New Challenges, New Opportunities

Editors: Manuela Lima, Amanda Ramos & Cristina Santos

eISSN (Online): 2405-7711

ISSN (Print): 2405-7703

eISBN (Online): 978-1-68108-385-8

ISBN (Print): 978-1-68108-386-5

Published by Bentham Science Publishers – Sharjah, UAE. All Rights Reserved.

BENTHAM SCIENCE PUBLISHERS LTD.

End User License Agreement (for non-institutional, personal use)

This is an agreement between you and Bentham Science Publishers Ltd. Please read this License Agreement carefully before using the ebook/echapter/ejournal (**"Work"**). Your use of the Work constitutes your agreement to the terms and conditions set forth in this License Agreement. If you do not agree to these terms and conditions then you should not use the Work.

Bentham Science Publishers agrees to grant you a non-exclusive, non-transferable limited license to use the Work subject to and in accordance with the following terms and conditions. This License Agreement is for non-library, personal use only. For a library / institutional / multi user license in respect of the Work, please contact: permission@benthamscience.org.

Usage Rules:

1. All rights reserved: The Work is the subject of copyright and Bentham Science Publishers either owns the Work (and the copyright in it) or is licensed to distribute the Work. You shall not copy, reproduce, modify, remove, delete, augment, add to, publish, transmit, sell, resell, create derivative works from, or in any way exploit the Work or make the Work available for others to do any of the same, in any form or by any means, in whole or in part, in each case without the prior written permission of Bentham Science Publishers, unless stated otherwise in this License Agreement.
2. You may download a copy of the Work on one occasion to one personal computer (including tablet, laptop, desktop, or other such devices). You may make one back-up copy of the Work to avoid losing it. The following DRM (Digital Rights Management) policy may also be applicable to the Work at Bentham Science Publishers' election, acting in its sole discretion:

- 25 'copy' commands can be executed every 7 days in respect of the Work. The text selected for copying cannot extend to more than a single page. Each time a text 'copy' command is executed, irrespective of whether the text selection is made from within one page or from separate pages, it will be considered as a separate / individual 'copy' command.
- 25 pages only from the Work can be printed every 7 days.

3. The unauthorised use or distribution of copyrighted or other proprietary content is illegal and could subject you to liability for substantial money damages. You will be liable for any damage resulting from your misuse of the Work or any violation of this License Agreement, including any infringement by you of copyrights or proprietary rights.

Disclaimer:

Bentham Science Publishers does not guarantee that the information in the Work is error-free, or warrant that it will meet your requirements or that access to the Work will be uninterrupted or error-free. The Work is provided "as is" without warranty of any kind, either express or implied or statutory, including, without limitation, implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the results and performance of the Work is assumed by you. No responsibility is assumed by Bentham Science Publishers, its staff, editors and/or authors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products instruction, advertisements or ideas contained in the Work.

Limitation of Liability:

In no event will Bentham Science Publishers, its staff, editors and/or authors, be liable for any damages, including, without limitation, special, incidental and/or consequential damages and/or damages for lost data and/or profits arising out of (whether directly or indirectly) the use or inability to use the Work. The entire liability of Bentham Science Publishers shall be limited to the amount actually paid by you for the Work.

General:

1. Any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims) will be governed by and construed in accordance with the laws of the U.A.E. as applied in the Emirate of Dubai. Each party agrees that the courts of the Emirate of Dubai shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims).
2. Your rights under this License Agreement will automatically terminate without notice and without the need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.
3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Ltd.
Executive Suite Y - 2
PO Box 7917, Saif Zone
Sharjah, U.A.E.
Email: subscriptions@benthamscience.org

CONTENTS

FOREWORD

The Genome science encompasses many scientific areas and is essential to advancing knowledge of different disciplines as Evolutionary Biology, Genetics, Medicine and Biological Anthropology, among others. With respect to this last discipline, the understanding of our place in nature has benefited in the recent years of the Genomic research that is being essential in the analysis of the biological origin of our species, and to measure in a more precise way human biological variation. The main genome projects have had a great impact in Biological Anthropology, and the anthropological studies are benefiting more and more from the genomic data. The developments in genome bioinformatics and computational biology have helped make the advances possible in the field of Anthropological Genetics, as well as in Forensic Anthropology and the population studies. Briefly, the current genomic revolution constitutes a turning point in our understanding of human evolution, and a fascinating insight into what can be revealed from the study of genomes.

The book deals with the genomic approaches in Biological Anthropology and how the DNA markers can provide insight into the processes of evolution and human variability; in addition, explanations of the technological developments and how they affect the fields of Forensic sciences and population studies are shown, alongside the methods of field investigations and their contribution to the Molecular Anthropology. But in turn, this volume reveals how the modern Anthropology can contribute to redefine the ways we have come to understand the genetic issues.

This volume, written by experts in their respective field, provides a rigorous overview about a subject so promising like Genomics and its present and future applications in the study of Human Biology. Through the different chapters authors are providing essential information for the area of Biological Anthropology that could change the traditional vision in this field and contribute to resolve numerous anthropological questions that have not yet been answered.

Dr. Esther Rebato
Department of Genetics, Physical Anthropology and Animal Physiology
University of the Basque Country (UPV-EHU)
Bilbao
Spain
E-mail: esther.rebato@ehu.es

PREFACE

Molecular methodologies have been routinely applied in Biological Anthropology to increase our understanding of human diversity and to elucidate the associations within and between human populations, as well as our evolutionary relationships with nonhuman primates. Since the first molecular studies, which date back to the late 1960s, molecular methodologies have been rapidly applied to investigations in the area of Biological Anthropology. The development of the Human Genome Project, concluded in 2004, resulted in the availability of a complete human reference sequence; combined with very important advances in sequencing and bioinformatics technologies, leading to other genome-scale projects.

In Chapter 1, Santos and co-authors present a general characterization of the main genome projects with potential impact in the field of Biological Anthropology, providing examples of questions to which genomic data can now successfully be called upon.

The emergence of genomics has contributed to the availability of databases containing large amounts of information, leading also to the implementation of new mathematical/ bioinformatics methods, which have undergone a major expansion in the recent years. The analysis of both nuclear and mitochondrial human genomes received new tools of analysis, providing information that now needs to be conciliated with previous classic genetic evidences. Discrepancies between the results obtained with the use of such methodological innovations and the most established methodologies thus constitute a challenge, which anthropologists need to resolve. The transition from traditional approaches to massively parallel sequencing or next-generation sequencing of the mitochondrial DNA (mtDNA) is discussed in Chapter 2, by Marques and collaborators; these authors highlight the need for the development and validation of new routine procedures and optimization of laboratorial protocols.

The expanding amount of data has made it possible to address several important questions in a molecular evolutionary context. In Chapter 3, Ramos and co-workers analyse the contribution of germinal *versus* somatic heteroplasmy, discussing its impact on aging and health.

Besides mtDNA, genomics has also been impacting the analysis of yet another monoparental system, the Y chromosome. Because a major goal of Biological Anthropology is to date events related with the present day populations (such as major migratory waves), mutation rate estimations for the Y chromosome are pertinent. In Chapter 4, Francalaci and collaborators address alternative methodologies which use genomic data to estimate such rates.

Information derived from genome projects, namely from the HapMap project is, as previously referred, having a tremendous impact on providing in-depth insights into the genetic makeup of human populations, namely isolated populations, who are privileged targets of gene-finding studies of both monogenic and multifactorial diseases. In Chapter 5, the characteristics of genetically isolated populations are addressed by Lima and the potential impact of genomic data on gene finding efforts is discussed.

In Chapter 6, Fatjó-Vilas & Arias further address the imports of the genomic era into discoveries concerning the aetiology of multifactorial diseases, focusing on ecogenetics, an area which studies the relationship between genetic and environmental factors, looking for gene-environment interactions.

Ancient DNA has also entered the genomic era; a historical overview of the advances and constraints in the field of ancient DNA analysis is provided by Simon & Malgosa that also discuss, in Chapter 8, the pitfalls of ancient DNA analysis and the strategies to circumvent them. The advances achieved by paleogenetics are also acknowledged by these authors, in Chapter 9.

Human genomics has been impacting several areas, and Forensic Anthropology is no exception. In Chapter 10, the interrelation between Forensic Anthropology and Forensic Genetics is highlighted, arguing that recent genomic tools have the potential to efficiently resolve questions left unanswered by genetics. Genomic resolution is gradually occupying an important place in Biological Anthropology; the process has been relatively simple in several sub-areas of Anthropology, such as population genetics, whereas in other sub-areas, such as Forensic Anthropology, several issues need to be addressed before a routine application of genomic data can be considered.

Initially tailored towards research of human health and disease, genomics has already provided important attention to human origins and variation studies; yet, and although paleogenomics has been dramatically impacting anthropology, genomics as an anthropological subject is still in its infancy and more work has yet to be done.

Dr. Manuela Lima
Department of Biology, University of the Azores
Ponta Delgada, Portugal

Dr. Amanda Ramos & Dr. Cristina Santos
Universitat Autònoma de Barcelona, Cerdanyola del Vallès
Barcelona, Spain

List of Contributors

Amanda Ramos

Unitat Antropologia Biològica, Department BABVE, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain
Department of Biology, University of the Azores, Ponta Delgada, Portugal
Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
Institute for Molecular and Cell Biology (IBMC), University of Porto, Porto, Portugal

Ana Goios

i3S-Instituto de Investigação e Inovação em Saúde/IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

Antonella Useli

Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy

Assumpció Malgosa Morera

Departament de Biologia Animal, Biologia Vegetal i Ecologia, Unitat d'Antropologia Biològica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Bárbara Arias

Departament de Biologia Animal, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain
Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain
Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain

Cristina Santos

Unitat Antropologia Biològica, Department BABVE, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

Daria Sanna

Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy

Eugénia Cunha

Department of Life Sciences, Center of Functional Ecology, University of Coimbra, Coimbra, Portugal

Luis Alvarez

i3S-Instituto de Investigação e Inovação em Saúde/IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

Mafalda Raposo

Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal
Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, Porto, Portugal

Manuela Lima

Department of Biology, University of the Azores, Ponta Delgada, Portugal
Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
Institute for Molecular and Cell Biology (IBMC), University of Porto, Porto, Portugal

Marc Simón Martínez

Departament de Biologia Animal, Biologia Vegetal i Ecologia, Unitat d'Antropologia Biològica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Mar Fatjó-Vilas

FIDMAG Germanes Hospitalàries Research Foundation, Sant Boi de Llobregat, Barcelona, Spain
Departament de Biologia Animal, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain
Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain
Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain

Maria Pilar Aluja

Unitat Antropologia Biològica, Department BABVE, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

Paolo Francalacci

Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy

Sofia L. Marques

i3S-Instituto de Investigação e Inovação em Saúde/IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

CHAPTER 1

Human Genomic Projects: Setting the Stage for Genome-Scale Anthropological Studies

Cristina Santos[1,*], **Mafalda Raposo**[2,3,4], **Amanda Ramos**[1,2,3,4] and **Manuela Lima**[2,3,4]

[1] *Unitat Antropologia Biològica, Department Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain*

[2] *Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal*

[3] *Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal*

[4] *Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal*

Abstract: With the development of the Human Genome Project, a complete reference sequence of the human genome became available. As new sequencing platforms and bioinformatics tools were continuously developed, sequencing costs were reduced. The emergence of several genome projects was grounded in such developments, allowing the scrutinizing, at the genome level, of the present human genetic variation, the analysis of extinct species (such as the Neanderthal) as well as the study of non-human primates. A general characterization of the main genome projects with potential impact in the field of Biological Anthropology is performed in this chapter. Examples of studies which profited from the genomic data produced are also provided. As high resolution studies are becoming affordable and faster, anthropologists around the world are being challenged to benefit and exploit the data being generated from the several international large-scale genomic projects. If they are able to meet this challenge, traditional questions of anthropological importance can be addressed in a new and much more efficient way.

Keywords: 1000 Genomes, ENCODE, Genographic, GTEx, HapMap, Human genome, Human genome projects, Neanderthal genome, Non-human primates genome projects, Roadmap Epigenomics.

[*] **Corresponding author Cristina Santos:** Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain; Tel/Fax: +34 93 5811503; +34 93 5811321; E-mail: cristina.santos@uab.cat

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

1.1. THE FOUNDATION OF GENOMICS: OVERVIEW OF THE HUMAN GENOME PROJECT

The emergence of DNA sequencing technology in the mid-1970s initiated a revolution in the field of Biology, anticipating the possibility of determining the sequence of the human genome and mapping all of its genes. The Human Genome Project (HGP) was officially launched in 1990 (www.genome.gov); in parallel with the technological and scientific endeavours, the discussion of the ethical, legal and social issues was also seen as a requirement for its proper development [1]. In 2001 a rough draft sequence representing a coverage of about 70% of the genome was published [2, 3]. The reference DNA sequence of the human genome, obtained almost entirely using the Sanger sequencing method, was completed in 2003 and published one year latter [4]; the information produced has been available, since the end of 2005, in public databases. Thereafter, human genome assembly data has been continuously improving, in order to produce an accurate consensus representation of the genome. The 2015 released primary assembly of the human genome accounts for 20296 coding genes, 25173 non-coding genes, 14424 pseudogenes and 198634 gene transcripts (http://www.ensembl.org/Homo_sapiens/Info/Annotation; GRCh38.p3 - Genome Reference Consortium Human Build 38).

The sequencing of the human genome was not the only scientific focus of the HGP; the value of sequencing the genomes of model organisms was also duly recognized. In this sense, *Escherichia coli, Saccharomyces cerevisiae* and *Mus musculus* constituted three of the several model species whose genome sequencing was considered as pilot study for the HGP [5].

With the development of new technologies [6], such as single nucleotide polymorphisms (SNPs) arrays (a technological tool that allows genotyping thousands of SNPs in a single experiment) and next generation sequencing (NGS) platforms, a complete reference sequence of the human genome became available; such progresses allowed new insights into human genetic diversity, in the context of several genomic projects (Table **1**). As new sequencing platforms and bioinformatics tools were continuously developed, sequencing costs were reduced, allowing the scrutinizing of human genetic variation at the whole genome scale. From such efforts emerged a huge amount of data which impacted not only the biomedical area, but also all aspects of biological anthropology.

Table 1. Most prominent genome projects with an impact in Biological Anthropology. The main goals as well as the first publication derived from these projects are reported.

Project	Duration	Main Goals
International HapMap Project	Phase I: 2003-2005 Phase II: 2005-2007 Phase III: 2007-2010	To determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain.
International HapMap Consortium. **The International HapMap Project**. Nature 2003; 426(6968): 789-96 [7]		
ENCODE	Pilot phase: 2004-2007 Expanded phase: 2007-2011	To provide a more biologically informative representation of the human genome by using high-throughput methods to identify and catalogue the functional elements encoded.
ENCODE Project Consortium. **The ENCODE (ENCyclopedia Of DNA Elements) Project**. Science 2004; 306(5696): 636-40 [8]		
THE GENOGRAPHIC PROJECT	2005	To answer fundamental questions about the origin of human populations and their migration paths.
1000 Genomes A Deep Catalog of Human Genetic Variation	Pilot phase: 2008-2009 Phase I: 2010-2011 Phase II: 2011	To sequence the genomes of at least 1,000 individuals from different populations around the world; to provide a comprehensive map of human genetic variation for future disease association studies and population genetics.
1000 Genomes Project Consortium, Abecasis GR, *et al.* **A map of human genome variation from population-scale sequencing**. Nature 2010; 467(7319): 1061-73 [9]		
ROADMAP epigenomics PROJECT	2010	To provide the scientific community ready access to a critical mass of high-quality epigenomic data for cells and tissues representative of normal human biology.
Bernstein BE, Stamatoyannopoulos JA, *et al.* **The NIH Roadmap Epigenomics Mapping Consortium**. Nat Biotechnol 2010; 28(10): 1045-8 [10]		
GTEx Portal	Pilot phase: 2013	To establish a resource database and associated tissue bank in which to study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues.
GTEx Consortium. **The Genotype-Tissue Expression (GTEx) project**. Nat Genet 2013; 45(6): 580-5 [11]		

1.2. HUMAN GENOMIC PROJECTS: GENERAL CHARACTERIZATION AND IMPACT ON BIOLOGICAL ANTHROPOLOGY

1.2.1. The HapMap Project

Initiated in 2003, the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov) aimed to create a reference map of common genetic variation [12] (Table **1**). The HapMap Consortium genotyped approximately 1.3 million SNPs (one common SNP every 5 kb across the genome) in 270 individuals from four geographically diverse populations. Samples used represented 30 trios of Yoruba people (Ibadan, Nigeria), 45 unrelated individuals from the Tokyo area (Japan) and 45 unrelated individuals from Beijing (China). Moreover 30 trios whose samples had been collected in 1980 from U.S. residents with northern and western European ancestry by the “Centre d'Etude du Polymorphisme Humain” (CEPH) were also included. The description of this resource was published in 2005 and was called HapMap Phase I [12]. In Phase II of the HapMap Project, 2.1 million SNPs were further successfully genotyped on the same individuals [13]. In Phase III the number of DNA samples was increased to 1301, originating from 11 distinct human populations.

Noteworthy, the HapMap Project allowed the description of the extent of linkage disequilibrium (LD) patterns across several populations; LD evaluates the co-occurrence, at a higher frequency than would be expected by chance, of a particular allele at a locus, with another allele at a distinct locus [14]. A major finding was that the human genome presents recombination hotspots and long segments of strong LD with limited haplotype diversity (haploblocks). Data produced by HapMap, concerning the haploblock structure of the human genome, has been used by the industry to develop cost-effective platforms that detect a limited number of haplotype-tagging variants (Tag SNPs), defined as SNPs that represent the allelic state of the markers in a region of high LD [15]. “Tag” SNPs have demonstrated to be portable between populations, having as a consequence the efficient capture of LD patterns, regardless of the population being studied [16]. Data gathered by the HapMap Phase I, II and III are freely available to the research community and have allowed the establishment of patterns of human variation and selection, at the genome level (for HapMap-related papers see http://hapmap.ncbi.nlm.nih.gov/publications.html.en).

Researchers in the field of biological anthropology have been making widely use of data from the HapMap project. As colleagues from other fields, biological anthropologists take advantage of the identification of Tag SNPs, and use the information derived to analyse many aspects related to human evolution and adaptation (see as example [17]). Data from different HapMap populations has also been used to create a map of selection across the human genome [18]. One interesting example was the study by Sabeti and collaborators [19] in which these authors used information concerning over 3 million polymorphisms from the International HapMap Project Phase 2 to identify evidences of positive selection [19]. Focused on 22 regions with strongest signals of positive selection, the previously referred authors found candidate genes related to skin pigmentation in Europe (*SLC24A5* and *SLC45A2*), to infection by the Lassa virus in West Africa (*LARGE* and *DMD*), and to the development of hair follicles in Asia (*EDAR* and *EDA2R*) [19].

1.2.2. The ENCyclopedia of DNA Elements (ENCODE) Project

Expression of protein-coding genes has been the initial focus of many studies during the last years. Notwithstanding, other functional genomic elements, such as non-protein-coding transcripts and transcription start sites, remained poorly understood. The development of ChIP-Seq methods, which combined chromatin immunoprecipitation (ChIP) assays with sequencing allowed to map, at the genome scale, binding sites for transcription factors, DNA binding proteins and histone modifications [20]. The mission of the ENCyclopedia Of DNA Elements (ENCODE) Project (Table **1**) was to discover and define the functional elements of the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns [21].

The ENCODE Project (www.encodeproject.org/) started by studying 1% of the human genome using several cell lines; results have revealed, amongst other aspects, that the human genome is pervasively transcribed, with several novel non-protein-coding transcripts being identified [22]. Other notable finding of this pilot project was that ~50% of the DNA elements identified were not conserved across mammals, thus raising more knowledge about mammalian evolution based on inter- and intra-species sequence comparisons [22]. Moreover, the reference human genome annotation derived from the ENCODE project (GENCODE-www.gencodegenes.org), aims to annotate all protein-coding genes, pseudogenes, and non-coding transcribed loci in the human genome (Fig. **1**) and to catalogue the products of transcription, including splice isoforms [23].

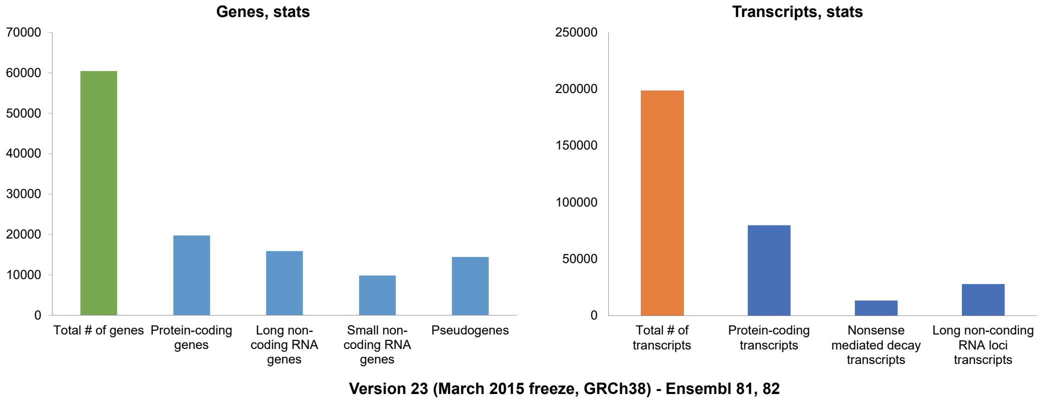


Fig. (1). Current human GENCODE release, version 23 statistics (http://www.gencodegenes.org).

Applications of the data derived from the ENCODE project are highly promising to the field of Biological Anthropology, since gene regulation mechanisms are thought to play an important role in the speciation and adaptation, namely on what concerns primates. As an example, Cain and collaborators [24] investigated the contribution of H3K4me3, a chromatin marker which is highly associated with gene repression [24]. The previous authors analysed regulatory differences in primates by looking at H3K4me3-associated genomic regions in lymphoblastoid cell lines (LCLs) from humans, chimpanzees, and rhesus macaques. Results from the previous study lead to the estimation that ~7% of gene expression differences between the LCLs of humans, chimpanzees, and rhesus macaques may be explained, at least partially, by changes in the status of H3K4me3 histone modifications [24].

1.2.3. National Geographic's Genographic Project

The National Geographic's Genographic Project (http://genographic. nationalgeographic.com/) was launched in 2005, describing itself as an ambitious attempt to answer fundamental questions about the origin of human populations and their migration paths (Table **1**), profiting from the collaboration of indigenous and more traditional communities. Additionally, the project was anonymous, without medical or financial purposes, hence proposing to solve several ethical issues associated with the commodification of genetic data, which were important obstacles to previous projects about human diversity, such as the pioneer Human Genome Diversity Project, led by Cavalli-Sforza [25].

Original samples used in the Genographic Project were collected in two ways. On one hand, the project comprised a consortium of ten scientific teams from around the world that were responsible for sample collection and analysis in their respective regions; on the other hand, the project promoted public participation so that anyone could take part on the project by purchasing a participation kit [26]. To date, more than 705,000 individuals have participated in the project and data obtained are available to the scientific community [27, 28].

The Genographic Project offered an opportunity for genetic based anthropology to prosper within the context of SNPs microarray technology, which had largely skipped this field [26]. As a discipline based on the study of humans and the respective populations, this has happened for three main reasons: 1) few groups of indigenous populations had been genotyped and studied; 2) comparisons between populations from different studies were not possible due to the low overlapping of different genotyping platforms; and 3) genotyping costs were prohibitively high for Anthropology studies, which rely on the analysis of large samples [26]. Furthermore, these arrays were enriched in trait- or disease-related markers and contained biased sets of pre-ascertained SNPs (which present higher minor allele frequencies than random SNPs), a fact that could induce errors in population genetics studies; in addition, markers used to access ancestry information were not accommodated [29]. The Genographic Project changed this avenue of research, opening the door to Anthropology since one of the major outcomes of the project was a cutting-edge new test kit - GenoChip - which has been specifically designed for studies in genetic anthropology [26]. This Chip originally included a unique collection of nearly 150000 markers that offer the richest ancestry-relevant information, excluding medical markers. The current version, Geno 2.0 Next Generation kit tests more than 750000 genetic markers. The GenoChip is a promising tool that should help to clarify the genetic relationships between archaic hominins such as Neanderthal and Denisovan (a recently discovered specie of *Homo* genus that coexisted with Neanderthals and modern humans), extinct humans and modern humans, as well as to provide a more detailed understanding of human migratory history [26]. In this context, GenoChip has been used in anthropological studies around the world; the study of indigenous populations of the Peruvian and Bolivian Altiplano [30] or the analysis of the genetic structure of India [31] are some of the examples.

1.2.4. 1000 Genomes Project

In recent years, data generated by large-scale NGS projects have promised a fuller evaluation of genetic variation of the nuclear genome. In contrast to SNP arrays, the study of the whole genome has allowed to construct, in a single step, a more comprehensive catalogue of different genetic variants such as SNPs, indels and structural variants [32]. One important contribution in this field was the 1000 Genomes Project [9, 33], launched in 2008 (Table **1**). The 1000 Genomes Project (www.1000genomes.org) was the first project which used NGS to obtain a deep characterization of human genome sequence variation, at a population scale [9, 33]. Moreover, the project also intended to develop methods and tools for sequence analysis that could be transferred to other sequencing projects. Both Pilot project and Phase 1 (which included the first 1092 samples) results were published and all data became available [9, 33]. Phase 2 increased the sample size to 1700 and improved methodological aspects. Phase 3 targeted more than 2500 samples from 26 different populations, including new African and South Asia samples.

Using 1000 Genomes data, a broad number of studies have been published (see for example [34] and [35]). Of particular relevance for its future application was the report of Willems and collaborators [35]; these authors reported the largest-scale analysis of human STR variation to date, based on the collection of information for nearly 700,000 short tandem repeat (STR) loci across more than 1000 individuals from Phase 1 of the 1000 Genomes Project [35]. Another important study with a direct application to population analysis was conducted by Gravel *et al.* [34], which reconstructed Native American migrations and presented new methods to estimate the allele frequencies in the Native American fraction of admixed American populations [34].

The bulk of work that has been developed based on the 1000 Genomes project ensures that variation identified by this project, particularly on what concerns rare variation, can provide determinant clues to understand adaptability of our specie, as well as susceptibility to disease.

1.2.5. The US National Institutes of Health (NIH) Roadmap Epigenomics Project

The National Institutes of Health (NIH) Roadmap Epigenomics project (www.roadmapepigenomics.org) was created to generate epigenomic maps in primary human tissue and cell types, including cells from a number of distinct brain regions and several types of immune cells [10] (Table **1**). Contrarily to the use of cell lines in the ENCODE project, which was considered as a limitation, the NIH Roadmap Epigenomics Project will provide epigenomes that are more physiologically representative of human cell types [36]. The analysis of the epigenomic maps, including DNA methylation, histone modifications and related chromatin features, are intended to be a reference for genomic and epigenomic studies in which the factors that underlie human development, diversity and disease are studied [10].

Although differences in the patterns of DNA methylation between closely related species have been reported, the extent to which such differences contribute to species-specific phenotypes has not been fully explored; these DNA methylation differences might represent an important molecular mechanism driving gene-expression divergence, namely between human and chimpanzee brains [37]. In fact, it has been postulated that epigenomic modifications rather than genetic changes could underlie the evolution of human-specific traits. To test this hypothesis, Schneider and collaborators (2014), compared DNA methylation patterns of *CNTNAP2* (which has been related with speech and language development) in 6 human and 5 chimpanzee cortices, finding significant methylation differences between them [38]. Moreover, the transcript variant CNTNAP2-201 was 1.6-fold upregulated in human cortex, compared with the chimpanzee [38]. These results suggests a role for *CNTNAP2* fine-regulation in human-specific language and communication traits after the human-chimpanzee splitting [38], demonstrating the importance that epigenetic mechanisms might have in the explanation of human lineage specific traits and reinforcing the pertinence of prosecuting the Roadmap Epigenomics project.

1.2.6. Genotype-Tissue Expression (GTEx) Project

The Genotype-Tissue Expression (GTEx) project (www.genome.gov/GTEx) aims to study human gene expression and regulation as well as its relationship to genetic variation, in multiple reference tissues (The GTEx Consortium, 2013) (Table **1**). The GTEx project began with a pilot phase (which lasted 2.5-years) to test the feasibility of establishing a rapid autopsy program that would yield high-quality nucleic acids and robust gene expression measurements from 47 different tissues, donated by 190 individuals (1814 total samples) [11]. After the donor enrolment and the RNA quality and expression quantitative trait locus (eQTL) findings were established, the project was scaled up to include approximately 900 post-mortem donors by the end of 2015 [11]. In 2015, the GTEx Consortium reported the analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, providing a landscape of gene expression in distinct tissues, cataloguing thousands of tissue-specific and shared regulatory eQTL variants, describing complex network relationships, and identifying signals from genome-wide association studies explained by eQTLs [39].

As it was previously underlined, the hypothesis of differences in gene regulation playing an important role in speciation and adaptation has been raised over 30 years ago. Several studies performed prior to the emergence of the GTEx project have demonstrated this to be a valid hypothesis, reinforcing the importance of developing this project. An example, the work of Gilad and collaborators [40] can be referred. The previously cited authors developed a study in which steady-state messenger RNA levels in liver tissues within and between humans, chimpanzees, orangutans and rhesus macaques were compared. A number of genes with similar expression levels among non-human primates but displaying significantly elevated or reduced expression in the human lineage were found, pinpointing to the action of directional selection [40].

1.3. THE NEANDERTHAL GENOME PROJECT: AN IN-DEPTH JOURNEY INTO OUR PAST

Neanderthals were first recognized as a distinct group of hominids from fossil remains discovered 150 years ago at Feldhofer in Neander Valley, outside Düsseldorf, Germany (revised in [41]). Subsequent Neanderthal findings in Europe and western Asia showed that fossils with Neanderthal traits appear in the fossil record of Europe and western Asia about 400,000 years ago and disappeared about 40,000 years ago [42].

The first Neanderthal DNA sequence was obtained by Krings and collaborators in 1997 [43]. Using a PCR based approach these authors recovered the mitochondrial DNA (mtDNA) hypervariable region 1 from the Feldhofer cave (Germany) Neanderthal specimen. By comparing the retrieved sequence against worldwide present-day human mtDNA sequences, they concluded that Neanderthals were a sister group to anatomically modern humans, providing no evidence of interbreeding between the two [43]. During the 15 years following this publication, other mtDNA Neanderthal sequences from different sites were successfully amplified with the same technical approach (for a review on the subject see [44]). A common observation of all these studies was that Neanderthal mtDNA sequences were similar to each other (suggesting general low diversity) and distinct to any reported sequences of modern human mtDNA.

With the introduction of NGS technologies to the field of ancient DNA (aDNA), it was possible for the first time to retrieve complete mitochondrial genomes [45, 46]. In this context, a multidisciplinary research consortium, led by Svante Pääbo at the Max Planck Institute for Evolutionary Anthropology, initiated in 2006 the Neanderthal genome project [47]. The aims of the Project, by sequencing the Neanderthal genome, were: a) to determine the sequences of the 3 billion base pairs that make up Neanderthal DNA; b) to provide a catalogue of differences between the human and Neanderthal genome; c) to determine what contribution, if any, Neanderthals made to modern human variation; d) to store this information in databases for public use; and e) to develop tools for ancient DNA data analysis. By 2010 the team reported the sequencing of an initial draft of the Neanderthal genome [45]. The sequence was generated from several Neanderthal fossils from Croatia, Germany, Spain and Russia, using high-throughput sequencing technologies. More recently, a high-quality Neanderthal genome sequence was generated from a toe bone discovered, in 2010, in Denisova Cave in southern Siberia [48]. Results indicated that Neanderthals are more closely related to modern humans outside Africa [41] and several gene flow events appear to have occurred among Neanderthals, Denisovans and early modern humans. Thus, even at low magnitude, interbreeding occurred among different hominin groups that coexist in the Late Pleistocene [43].

1.4. GENOMICS OF OUR RELATIVES: THE NON-HUMAN PRIMATE GENOME PROJECTS

With the accomplishment of the HGP, the scientific community recognized that the knowledge of our genome sequence should be followed by understanding the functions of human genes. In this sense, many efforts, since the conclusion of the HGP, using different approaches, have been undertaken to better understand the significance of human variation. Many researchers immediately advocated the need for a primate genome project (see for example [49]). Citing McConkey and Varki (2000, 1295) [49] "We cannot fully understand human genome function until we have identified genetic features that underlie uniquely human anatomical, physiological, behavioural and cognitive characteristics. To identify uniquely human aspects of gene structure and expression requires comparative data on related species. The mouse genome project will help, but analysis of rodent genomes can never tell us why we are not apes".

Table 2. Nonhuman primate genome projects. For each specie being sequenced, the status of the project as well as the accession to Genbank assembly are reported.

Common Name	Species Name	Status of the Project	Genome Assembly	Genbank Assembly Accession (Last Version)	Web Page
Chimpanzee	*Pan troglodytes*	Published [50]	Yes	GCA_000001515.4	http://genome.wustl.edu/genomes/detail/pan-troglodytes/ https://www.hgsc.bcm.edu/non-human-primates/chimpanzee-genome-project
Orang-utan	*Pongo abelii*	Published [51]	Yes	GCF_000001545.4	http://genome.wustl.edu/genomes/detail/pongo-abelii/ https://www.hgsc.bcm.edu/non-human-primates/orangutan-genome-project
Indian rhesus macaque	*Macaca mulatta*	Published [52]	Yes	GCA_000002255.2	http://genome.wustl.edu/genomes/detail/macaca-mulatta/ https://www.hgsc.bcm.edu/non-human-primates/rhesus-monkey-genome-project
Gorilla	*Gorilla gorilla*	Published [53]	Yes	GCA_000151905.3	http://www.sanger.ac.uk/resources/downloads/gorilla/
Bonobo	*Pan paniscus*	Published [54]	Yes	GCA_000258655.2	http://www.eva.mpg.de/bonobo-genome/index.html
Aye-aye	*Daubentonia madagascarensis*	Published [55]	Yes	GCA_000241425.1	NA
Chinese rhesus macaque	*Macaca mulatta*	Published [56]	Yes	GCA_000230795.1	NA
Vietnamese cynomolgus macaque	*Macaca fascicularis*	Published [56]	Yes	GCA_000364345.1	NA

(Table 2) contd.....

Common Name	Species Name	Status of the Project	Genome Assembly	Genbank Assembly Accession (Last Version)	Web Page
Gibbon	*Nomascus leucogenys*	Published [57]	Yes	GCA_000146795.3	https://www.hgsc.bcm.edu/non-human-primates/gibbon-genome-project
Marmoset	*Callithrix jacchus*	Published [58]	Yes	GCA_000004665.1	https://www.hgsc.bcm.edu/non-human-primates/marmoset-genome-project
Baboon	*Papio anubis*	Assembly	Yes	GCA_000264685.1	https://www.hgsc.bcm.edu/non-human-primates/baboon-genome-project
Black and white colobus	*Colobus angolensis*	Assembly	Yes	GCA_000951035.1	https://www.hgsc.bcm.edu/non-human-primates/black-and-white-colobus-genome-project
Drill	*Mandrillus leucophaeus*	Assembly	Yes	GCA_000951045.1	https://www.hgsc.bcm.edu/non-human-primates/drill-genome-project
Mouse lemur	*Microcebus murinus*	Assembly	Yes	GCA_000165445.2	https://www.hgsc.bcm.edu/non-human-primates/mouse-lemur-genome-project
Pigtail macaque	*Macaca nemestrina*	Assembly	Yes	GCA_000956065.1	https://www.hgsc.bcm.edu/non-human-primates
Sifaka	*Propithecus coquereli*	Assembly	Yes	GCA_000956105.1	https://www.hgsc.bcm.edu/non-human-primates/sifaka-lemur-genome-project
Sooty mangabey	*Cercocebus atys*	Assembly	Yes	GCA_000955945.1	https://www.hgsc.bcm.edu/non-human-primates/sooty-mangabey-genome-project
Squirrel monkey	*Saimiri boliviensis*	Assembly	Yes	GCA_000235385.1	http://moma.ki.au.dk/genome-mirror/cgi-bin/hgGateway?hgsid=7717591_aaW27IEP9rldm059XohU4RIgFP5m

(Table 2) contd.....

Common Name	Species Name	Status of the Project	Genome Assembly	Genbank Assembly Accession (Last Version)	Web Page
African green monkey	*Chlorocebus aethiops*	Analysis in progress	No	-	http://genome.wustl.edu/projects/detail/vervet-reference-genome-project/
Owl monkey	*Aotus nancymaae*	Sequencing in progress	No	-	https://www.hgsc.bcm.edu/non-human-primates/owl-monkey-genome-project

White papers gave high priority in early 2003 for the sequencing of both Chimpanzee (www.hgsc.bcm.edu/sites/default/files/documents/ChimpGenome2.pdf) and Rhesus macaque (*Macaca mulatta*) (www.genome.gov/Pages/Research/Sequencing/SeqProposals/RhesusMacaqueSEQ021203.pdf) genomes (Table **2**). The rational for the priority given to these two organisms resides in their unique placement in the evolutionary tree, relatively to humans as well as in the interest of the Rhesus macaque as a biomedical research model. Both Chimpanzee and Rhesus macaque genomes were analysed using exclusively Sanger sequencing methods. Thus, these projects implied considerable cost and effort. Since these initial initiatives, several institutions have been enrolled in sequencing genomes from representative species from most of the major branches of the phylogeny: hominoids, Old World monkeys, New World monkeys and lemurs (Table **2**). Some of them were initiated when Sanger sequencing was the only option, but advances in NGS allowed an explosion in the number of primate genomes being analysed, with a main goal to obtain a reference sequence for different species. Current sequencing efforts also all include diversity panels, investigating within-species variation or closely related species or subspecies. An extended review of primate genome projects can be found in Rogers and Gibbs (2014) [59].

Recently, two projects of great relevance were launched, the "Great Ape Genome Project" [60] and the Nonhuman Primate Reference Transcriptome Resource (NHPRTR) [61]. The first one, finished in 2013 [60], provided a high coverage sequence of a total of 79 wild- and captive-born individuals representing all six great ape species and 10 subspecies across the African continent and Southeast Asia. The NHPRTR was initiated in mid-2010. The concept was to develop a nonhuman primate reference transcriptome resource, consisting of deep sequencing complete transcriptomes from multiple nonhuman primate species.

CONCLUDING REMARKS

Following the HGP, several genome projects, summarized in this chapter were developed. Data from these projects could be joined to produce synergistic interactions. An example of how such interactions are possible is the study by Montgomery and collaborators [62] in which transcriptome data from 60 HapMap European individuals was analysed. Data was further integrated with SNP information from the HapMap Phase III, leading to the identification of novel eQTLs and sequence variants responsible for alternative splicing [62]. The link between GTEx and ENCODE projects can also be pinpointed; complementing eQTL results with epigenomic maps, and linking this information with ENCODE data will allow new knowledge on gene-regulatory mechanisms and networks across multiple tissues [11].

In conclusion, the genomics era has come to stay; high resolution studies are going to become cheaper and faster and they are now a reality that is slowly reaching anthropological genetic laboratories. Despite the potential of these new technologies and methodologies, it must be noted that they require training in novel laboratorial and bioinformatics methods, representing an important challenge to the field. If anthropologists around the world, and not only a small subset of researchers integrated on large-scale genotyping projects, are able to meet this challenge, traditional questions of anthropological importance can be addressed in new and much more efficient ways. Moreover, anthropologists will fully take benefit and exploit the data that are being generated with the international large-scale genomic projects.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Bharadwaj M. Looking back, looking beyond: revisiting the ethics of genome generation. J Biosci 2006; 31(1): 167-76.
[http://dx.doi.org/10.1007/BF02705245] [PMID: 16595885]

[2] Lander ES, Linton LM, Birren B, *et al.* International human genome sequencing consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409(6822): 860-921.
[http://dx.doi.org/10.1038/35057062] [PMID: 11237011]

[3] Venter JC, Adams MD, Myers EW, *et al.* The sequence of the human genome. Science 2001; 291(5507): 1304-51.
[http://dx.doi.org/10.1126/science.1058040] [PMID: 11181995]

[4] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431(7011): 931-45.
[http://dx.doi.org/10.1038/nature03001] [PMID: 15496913]

[5] National Human Genome Research Institute (NHGRI). Education: All about the human genome project (HGP). Available from: www.genome.gov/10001772#al-4 2015.

[6] Veeramah KR, Hammer MF. The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 2014; 15(3): 149-62.
[http://dx.doi.org/10.1038/nrg3625] [PMID: 24492235]

[7] International HapMap Consortium. The International HapMap Project. Nature 2003; 426(6968): 789-96.
[http://dx.doi.org/10.1038/nature02168] [PMID: 14685227]

[8] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004; 306(5696): 636-40.
[http://dx.doi.org/10.1126/science.1105136] [PMID: 15499007]

[9] Abecasis GR, Altshuler D, Brooks LD. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319): 1061-73.

[10] Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 2010; 28(10): 1045-8.
[http://dx.doi.org/10.1038/nbt1010-1045] [PMID: 20944595]

[11] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013; 45(6): 580-5.
[http://dx.doi.org/10.1038/ng.2653] [PMID: 23715323]

[12] International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437(7063): 1299-320.
[http://dx.doi.org/10.1038/nature04226] [PMID: 16255080]

[13] Frazer KA, Ballinger DG, Cox DR, *et al.* International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449(7164): 851-61.
[http://dx.doi.org/10.1038/nature06258] [PMID: 17943122]

[14] Rahim NG, Harismendy O, Topol EJ, Frazer KA. Genetic determinants of phenotypic diversity in humans. Genome Biol 2008; 9(4): 215.
[http://dx.doi.org/10.1186/gb-2008-9-4-215] [PMID: 18439327]

[15] Palotie A, Widén E, Ripatti S. From genetic discovery to future personalized health research. N Biotechnol 2013; 30(3): 291-5.
[http://dx.doi.org/10.1016/j.nbt.2012.11.013] [PMID: 23165095]

[16] González-Neira A, Ke X, Lao O, *et al.* The portability of tagSNPs across populations: a worldwide survey. Genome Res 2006; 16(3): 323-30.
[http://dx.doi.org/10.1101/gr.4138406] [PMID: 16467560]

[17] Jeong C, Alkorta-Aranburu G, Basnyat B, *et al.* Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun 2014; 5: 3281.
[http://dx.doi.org/10.1038/ncomms4281] [PMID: 24513612]

[18] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol 2006; 4(3): e72.
[http://dx.doi.org/10.1371/journal.pbio.0040072] [PMID: 16494531]

[19] Sabeti PC, Varilly P, Fry B, *et al.* International HapMap Consortium. Genome-wide detection and characterization of positive selection in human populations. Nature 2007; 449(7164): 913-8.
[http://dx.doi.org/10.1038/nature06250] [PMID: 17943131]

[20] Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku C-S. Human genetics and genomics a decade after the release of the draft sequence of the human genome. Hum Genomics 2011; 5(6): 577-622.
[http://dx.doi.org/10.1186/1479-7364-5-6-577] [PMID: 22155605]

[21] ENCODE Project Consortium. A users guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 2011; 9(4): e1001046.
[http://dx.doi.org/10.1371/journal.pbio.1001046] [PMID: 21526222]

[22] Birney E, Stamatoyannopoulos JA, Dutta A, *et al.* ENCODE project consortium, NISC comparative sequencing program, baylor college of medicine human genome sequencing center, Washington university genome sequencing center, broad institute, childrens hospital Oakland research institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447(7146): 799-816.
[http://dx.doi.org/10.1038/nature05874] [PMID: 17571346]

[23] Harrow J, Frankish A, Gonzalez JM, *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012; 22(9): 1760-74.
[http://dx.doi.org/10.1101/gr.135350.111] [PMID: 22955987]

[24] Cain CE, Blekhman R, Marioni JC, Gilad Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics 2011; 187(4): 1225-34.
[http://dx.doi.org/10.1534/genetics.110.126177] [PMID: 21321133]

[25] Cavalli-Sforza LL. The human genome diversity project: past, present and future. Nat Rev Genet 2005; 6(4): 333-40.
[http://dx.doi.org/10.1038/nrg1596] [PMID: 15803201]

[26] Elhaik E, Greenspan E, Staats S, *et al.* Genographic Consortium. The GenoChip: a new tool for genetic anthropology. Genome Biol Evol 2013; 5(5): 1021-31.
[http://dx.doi.org/10.1093/gbe/evt066] [PMID: 23666864]

[27] Behar DM, Harmant C, Manry J, *et al.* Genographic Consortium. The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. Am J Hum Genet 2012; 90(3): 486-93.
[http://dx.doi.org/10.1016/j.ajhg.2012.01.002] [PMID: 22365151]

[28] Martinez-Cruz B, Ioana M, Calafell F, *et al.* Genographic Consortium. Y-chromosome analysis in individuals bearing the Basarab name of the first dynasty of Wallachian kings. PLoS One 2012; 7(7): e41803.
[http://dx.doi.org/10.1371/journal.pone.0041803] [PMID: 22848614]

[29] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays 2013; 35(9): 780-6.

[http://dx.doi.org/10.1002/bies.201300014] [PMID: 23836388]

[30] Sandoval JR, Lacerda DR, Jota MS, *et al.* Genographic Project Consortium. The genetic history of indigenous populations of the Peruvian and Bolivian Altiplano: the legacy of the Uros. PLoS One 2013; 8(9): e73006.
[http://dx.doi.org/10.1371/journal.pone.0073006] [PMID: 24039843]

[31] ArunKumar G, Tatarinova TV, Duty J, *et al.* Genographic consortium. Genome-wide signatures of male-mediated migration shaping the Indian gene pool. J Hum Genet 2015; 60(9): 493-9.
[http://dx.doi.org/10.1038/jhg.2015.51] [PMID: 25994871]

[32] Stapley J, Reger J, Feulner PG, *et al.* Adaptation genomics: the next generation. Trends Ecol Evol (Amst) 2010; 25(12): 705-12.
[http://dx.doi.org/10.1016/j.tree.2010.09.002] [PMID: 20952088]

[33] Abecasis GR, Auton A, *et al.* 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491(7422): 56-65.
[http://dx.doi.org/10.1038/nature11632]

[34] Gravel S, Zakharia F, Moreno-Estrada A, *et al.* 1000 Genomes project. Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet 2013; 9(12): e1004023.
[http://dx.doi.org/10.1371/journal.pgen.1004023] [PMID: 24385924]

[35] Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. 1000 Genomes Project Consortium. The landscape of human STR variation. Genome Res 2014; 24(11): 1894-904.
[http://dx.doi.org/10.1101/gr.177774.114] [PMID: 25135957]

[36] Siggens L, Ekwall K. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. J Intern Med 2014; 276(3): 201-14.
[http://dx.doi.org/10.1111/joim.12231] [PMID: 24605849]

[37] Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am J Hum Genet 2012; 91(3): 455-65.
[http://dx.doi.org/10.1016/j.ajhg.2012.07.024] [PMID: 22922032]

[38] Schneider E, El Hajj N, Richter S, *et al.* Widespread differences in cortex DNA methylation of the language gene CNTNAP2 between humans and chimpanzees. Epigenetics 2014; 9(4): 533-45.
[http://dx.doi.org/10.4161/epi.27689] [PMID: 24434791]

[39] Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015; 348(6235): 648-60.
[http://dx.doi.org/10.1126/science.1262110] [PMID: 25954001]

[40] Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 2006; 440(7081): 242-5.
[http://dx.doi.org/10.1038/nature04559] [PMID: 16525476]

[41] Schmitz RW, Serre D, Bonani G, *et al.* The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proc Natl Acad Sci USA 2002; 99(20): 13342-7.
[http://dx.doi.org/10.1073/pnas.192464099] [PMID: 12232049]

[42] Higham T, Douka K, Wood R, *et al.* The timing and spatiotemporal patterning of Neanderthal disappearance. Nature 2014; 512(7514): 306-9.
[http://dx.doi.org/10.1038/nature13621] [PMID: 25143113]

[43] Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. Neandertal DNA sequences and the origin of modern humans. Cell 1997; 90(1): 19-30.
[http://dx.doi.org/10.1016/S0092-8674(00)80310-4] [PMID: 9230299]

[44] Sánchez-Quinto F, Lalueza-Fox C. Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. Philos Trans R Soc Lond B Biol Sci 2015; 370(1660): 20130374.
[http://dx.doi.org/10.1098/rstb.2013.0374] [PMID: 25487326]

[45] Green RE, Krause J, Briggs AW, *et al.* A draft sequence of the Neandertal genome. Science 2010; 328(5979): 710-22.
[http://dx.doi.org/10.1126/science.1188021] [PMID: 20448178]

[46] Briggs AW, Good JM, Green RE, *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009; 325(5938): 318-21.
[http://dx.doi.org/10.1126/science.1174462] [PMID: 19608918]

[47] The neanderthal genome project. Leipzig; Max planck institute for evolutionary anthropology. Department of evolutionary genetics, genome projects - neanderthal genome. Available from: http://www.eva.mpg.de/ neandertal/draft-neandertal-genome/about.html 2016.

[48] Prüfer K, Racimo F, Patterson N, *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 2014; 505(7481): 43-9.
[http://dx.doi.org/10.1038/nature12886] [PMID: 24352235]

[49] McConkey EH, Varki A. A primate genome project deserves high priority. Science 2000; 289(5483): 1295-6.
[http://dx.doi.org/10.1126/science.289.5483.1295b] [PMID: 10979852]

[50] Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005; 437(7055): 69-87.
[http://dx.doi.org/10.1038/nature04072] [PMID: 16136131]

[51] Locke DP, Hillier LW, Warren WC, *et al.* Comparative and demographic analysis of orang-utan genomes. Nature 2011; 469(7331): 529-33.
[http://dx.doi.org/10.1038/nature09687] [PMID: 21270892]

[52] Gibbs RA, Rogers J, Katze MG, *et al.* Rhesus macaque genome sequencing and analysis consortium. Evolutionary and biomedical insights from the rhesus macaque genome. Science 2007; 316(5822): 222-34.
[http://dx.doi.org/10.1126/science.1139247] [PMID: 17431167]

[53] Scally A, Dutheil JY, Hillier LW, *et al.* Insights into hominid evolution from the gorilla genome sequence. Nature 2012; 483(7388): 169-75.
[http://dx.doi.org/10.1038/nature10842] [PMID: 22398555]

[54] Prüfer K, Munch K, Hellmann I, *et al.* The bonobo genome compared with the chimpanzee and human genomes. Nature 2012; 486(7404): 527-31.

[http://dx.doi.org/10.1038/nature11128] [PMID: 22722832]

[55] Perry GH, Reeves D, Melsted P, *et al.* A genome sequence resource for the aye-aye (*Daubentonia madagascariensis*), a nocturnal lemur from Madagascar. Genome Biol Evol 2012; 4(2): 126-35. [http://dx.doi.org/10.1093/gbe/evr132] [PMID: 22155688]

[56] Yan G, Zhang G, Fang X, *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 2011; 29(11): 1019-23. [http://dx.doi.org/10.1038/nbt.1992] [PMID: 22002653]

[57] Carbone L, Harris RA, Gnerre S, *et al.* Gibbon genome and the fast karyotype evolution of small apes. Nature 2014; 513(7517): 195-201. [http://dx.doi.org/10.1038/nature13679] [PMID: 25209798]

[58] The common marmoset genome provides insight into primate biology and evolution. Nat Genet 2014; 46(8): 850-7. [http://dx.doi.org/10.1038/ng.3042] [PMID: 25038751]

[59] Rogers J, Gibbs RA. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet 2014; 15(5): 347-59. [http://dx.doi.org/10.1038/nrg3707] [PMID: 24709753]

[60] Prado-Martinez J, Sudmant PH, Kidd JM, *et al.* Great ape genetic diversity and population history. Nature 2013; 499(7459): 471-5. [http://dx.doi.org/10.1038/nature12228] [PMID: 23823723]

[61] Nonhuman Primate Reference Transcriptome Resource [homepage on the Internet]. 2014. [cited: 15th October 2015]. Available from: http://nhprtr.org/index.html

[62] Montgomery SB, Sammeth M, Gutierrez-Arcelus M, *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 2010; 464(7289): 773-7. [http://dx.doi.org/10.1038/nature08903] [PMID: 20220756]

CHAPTER 2

Complete Mitochondrial DNA through Massively Parallel Sequencing: Methodology and Applications

Sofia L. Marques, Ana Goios and **Luis Alvarez**[*]

i3S-Instituto de Investigação e Inovação em Saúde/IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

Abstract: Sanger sequencing has been the standard method for mtDNA typing, however, over the past years, new technologies are rapidly evolving to overcome the limitations of the Sanger biochemistry approach. Massively parallel sequencing (MPS) or next-generation sequencing (NGS) technologies allow the sequencing of the entire mitochondrial genome at once, and the simultaneous sequencing of a large number of different samples. The technologies and software tools are rapidly evolving, and at the moment, several MPS platforms and different sequencing strategies are available, with their inherent advantages and limitations. However, the transition from traditional approaches to MPS-based tests demands for the development and validation of new routine procedures and optimized laboratorial protocols. When properly validated these systems can be applied to a wide range of fields spanning from Forensic and Population genetics to clinical casework. In this chapter, we present an overview of the currently available MPS sequencing methodologies for mtDNA analysis and discuss the advantages and limitations for each of the different applications.

Keywords: Amplification strategies, Ancient mtDNA, Applications, Diagnostics, Forensic investigation, Illumina, Ion torrent, Long-range PCR, Miseq, Mitochondrial disorders, Mitogenome, MPS platforms, MtDNA, PGM, Phylogeography, Population genetics, Semiconductor, Sequencing, Sequencing by synthesis, Validation.

[*] **Corresponding author Luis Alvarez:** IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto Rua Júlio Amaral de Carvalho 45, Porto, Portugal; Tel: 00351 225570700; Fax: 00351 225570799; E-mail: lfernandez@ipatimup.pt

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

2.1. INTRODUCTION

Mammalian mitochondrial DNA (mtDNA) is a circular double-stranded molecule of ~16.6 kb that encodes for 13 protein-coding genes, 2 ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs), all essential in the process of oxidative phosphorylation (OXPHOS). Unlike nuclear genes, no introns or non-coding regions are scattered through the mitochondrial genome, except for the displacement loop (D-loop) or Control Region (CR) which is believed to play a regulatory role.

Although Sanger sequencing has been the standard method for mtDNA typing, over the past years new technologies have been rapidly evolving to overcome the methodology limitations of the Sanger biochemistry approach. Massively Parallel Sequencing (MPS) or Next-Generation Sequencing (NGS) technologies provide a wide-range coverage of the genome per sample, allowing the sequencing of the entire mitochondrial genome at once and the simultaneous sequence of a large number of different samples. The cost of DNA sequencing and laboratory time preparation are also significantly reduced. In fact, although significant hurdles remain in what comes to a straightforward implementation of MPS in routine research and diagnostic laboratories, when appropriately validated, this high-throughput sequencing technology has the potential to revolutionize prognostic/diagnostic routine casework and accelerate biomedical research.

In the following sections complete mtDNA sequencing methodologies and applications, as well as the current available MPS technologies strengths and weaknesses will be exposed. A particular attention will be given to compact MPS platforms, which are the most practical and simple systems, with a genome throughout and adequate sensibility for mitochondrial studies.

2.2. MTDNA CHARACTERISTICS AND APPLICATIONS

The mtDNA (or mitogenome) differs from the nuclear genome (nDNA) in several characteristics that make it a valuable genetic marker for a wide-range of areas, from Population and Forensic genetics to the diagnostic field:

i. mtDNA is present in much higher copy number per cell than nDNA, since each mitochondrion can contain several copies of mitogenomes and a single cell can contained hundreds of mitochondria [1].
ii. mtDNA is a maternally inherited haploid genome transmitted without recombination, leading to an effective population size (N_e) about one-quarter relative to nDNA [2].
iii. mtDNA mutation rate is higher than that observed for nDNA, and varies throughout the genome, despite being almost exclusively comprised by coding regions. This high mutation rate is largely due to the exposure of mtDNA to oxidative damage as result of the reactive oxygen species produced in oxidative phosphorylation [3, 4].
iv. mtDNA circular nature and the fact that is encapsulated in a double membrane-bound organelle make it also less susceptible to exonucleases that degrade DNA, increasing mtDNA survival rate [1].

The high copy number and the resistance to degradation are extremely useful in a forensic context when samples fail to yield successful nDNA profiles [5], on specific forensic scenarios when DNA needs to persist until forensic testing, or even when ancient samples are under study. Although complete mtDNA sequencing is not yet feasible in the context of routine forensic casework, it remains essential for quality control and databasing. Moreover, complete mtDNA can be relevant when the aim of DNA profiling is to maximize the discrimination power between individuals.

Because hundreds to thousands of copies of mtDNA molecules coexist in a single cell, tissue or individual, there may be states where different types of molecules (haplotypes) are present. This state is known as heteroplasmy and although it may occur at determined positions in healthy individuals, it is frequent in cases of mutations associated with mitochondrial disease. In fact, there is a heteroplasmy threshold, variable with the type of mutation, needed for the disease phenotype to show.

Most mtDNA mutations are accumulated due to their neutrality, resulting in populations of healthy individuals with shared mutations that can be grouped in maternal lineages/haplogroups. The low N_e renders the mtDNA molecule more sensitive to demographic alterations, and haplogrouping allows to trace back maternal lineages to clarify the demographic history of populations. In this context, the analysis of the complete mtDNA molecule has been used to increase the resolution of mtDNA phylogeny from a rougher global to a finer micro-geographic scale. However, several mtDNA mutations (point mutations and large-scale rearrangements) have been demonstrated to cause disease syndromes ([6], http://www.mitomap.org) with variable age of onset and clinical presentations, from a single affected organ to multiple tissue/organ dysfunctions and with symptoms being more critical in highly energy demanding tissues/organs such as brain and muscle [7]. Mitochondrial disorders can directly affect OXPHOS or other mitochondrial functions. Despite the high number of pathogenic mutations reported in the mtDNA, a large number of mitochondrial diseases are caused by nuclear gene mutations. More than 1300 nuclear genes are responsible for mitochondrial functions and structure [8] resulting in a high heterogeneity of mitochondrial diseases caused by those genes. In this way, an additional advantage of using MPS technologies in this field is the ability to sequence several genes, nuclear or mitochondrial, in a single test and diagnostic report.

2.3. AMPLIFICATION STRATEGIES

Currently there are several target-enrichment approaches available. Although some strategies avoid a PCR amplification – such as molecular inversion probes (MIPs) and hybrid techniques – PCR continues to be the strategy with higher levels of sensitivity, specificity, uniformity and reproducibility, being also compatible with any MPS technology [9]. PCR amplification can be laborious and expensive when applied to very long targeted regions (>30 Mb), which is not a concern in the relatively short mtDNA (~16,5Kb), and it may alter the proportion of the different mtDNA molecules in a heteroplasmic sample due to the exponential amplification effect. However, these disadvantages may be at least partially overcome as we discuss later on, and thus PCR continues to be the most employed enrichment strategy for routine mtDNA analysis and receives a special focus throughout this chapter.

The few non-coding segments in mtDNA are essentially located in the CR (Fig. **1**). Inside this region, some segments are highly conserved while others are variable, the so-called Hypervariable Regions (HVR-I; HVR-II and HVR-III). Until recently only these highly polymorphic regions of the CR were commonly analyzed for phylogeography or forensic routine casework. Standardly, each laboratory uses an amplification strategy that is mainly dependent on the location of the selected primers. However, the latest recommendations from the International Society for Forensic Genetics (ISFG) specify the need for an amplification, whenever possible, of the entire CR in a single amplicon and the use of both forward and reverse primers for sequencing [5, 10].

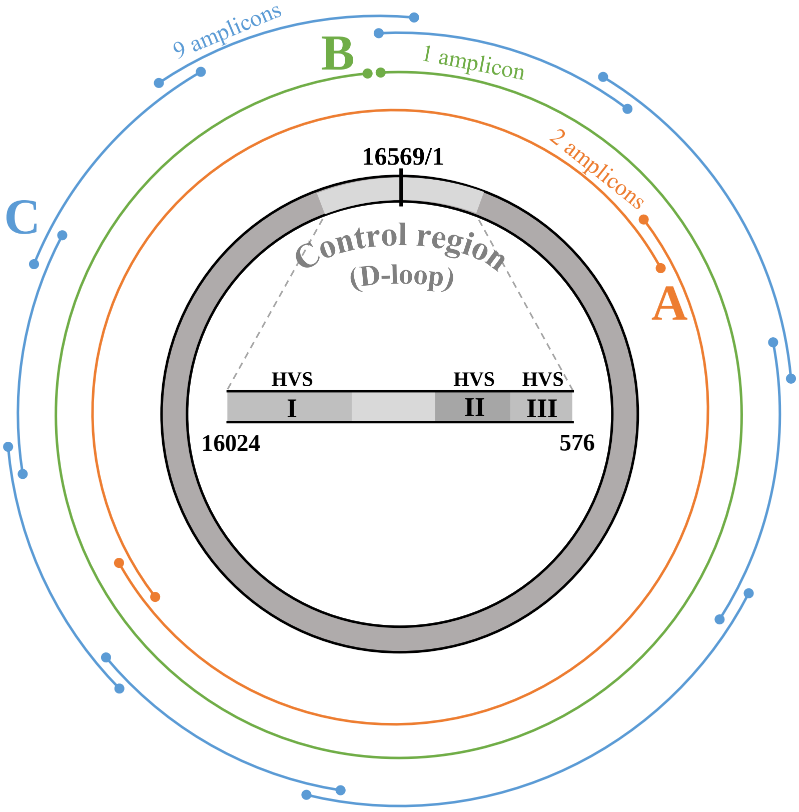


Fig. (1). Schematic representation of complete mtDNA amplification strategies. (**A-B**) LR-PCR, with two overlapping fragments and a single amplicon, respectively; (**C**) classical PCR, with nine overlapping fragments.

Regarding the complete mitogenomes amplification, commonly required for clinical diagnosis, there are several strategies available in the literature. These protocols differ mainly by the number and size of amplicons (Fig. **1**) and can be classified in two main categories: i) classical PCR approach, including a variable number of overlapping fragments [11, 12]; and ii) long-range PCR (LR-PCR) approach, with the amplification of two overlapping segments [13, 14] or even the complete molecule in a single amplicon [15, 16]. However, the allocation of the primers along the molecule is not a negligible issue, depending on the purpose of the analysis special attention is required, as for instance when the aim of the designed approach is to prevent the amplification of nDNA sequences of mitochondrial origin (NUMTs) [17]. All these amplification strategies can be applied for traditional Sanger sequencing approach but also for the use of MPS systems.

2.3.1. Long-range PCR

The LR-PCR approach is a fast and cost-effective tool for mtDNA amplification that simplifies laboratory routine, especially when combined with MPS sequencing platforms. It also offers an advantage on specificity issues, the amplification of large amplicon sizes or preferentially a single amplicon not only allows uniform mtDNA coverage but also avoids NUMT amplification [15]. Moreover, because only one fragment represents the complete mtDNA, the potential preferential amplification of one type of molecule in a heteroplasmic mixture is equalized throughout the molecule. Although this does not prevent an erroneous evaluation of the heteroplasmic levels, it at least allows for a homogeneous detection of the different variants along the molecule.

The use of combined enzymes is one of the most common strategies implemented to increase amplification fidelity in long-range amplifications. In fact, commercially available optimized polymerases can amplify up to 20 kb in a single PCR reaction; and high-fidelity and sensitive enzymes, with low error rates, successfully amplify low input DNA and GC-rich templates. Usually, the traditional *Thermus aquaticus* (*Taq*) DNA polymerase (5′→3′ polymerase activity) is complemented by the inclusion of another polymerase with proofreading activity. These proofreading enzymes have a 3′→ 5′ exonuclease activity that allow the repair of misincorporations and therefore let to a more accurate PCR. It is worth mentioning that proofreading enzymes alone have also a poor productivity resulting in low amplicon sizes, which justifies its use in combination with *Taq* DNA polymerases.

However, the wide range of available products can make the choice of the most appropriate enzyme in each particular case a difficult task. In a recent report, Jia and collaborators (2014) [18] tested the real performance of some commercially available polymerases. The obtained results clearly indicated that enzymes' performance varies greatly between manufacturers and few work properly without optimizing experimental conditions depending on the target/primers used. Manufacturers' websites generally provide tools to help researchers in the selection of the suitable enzyme for a particular study, having into account the characteristics of the DNA template used and the target regions to be amplified, as well as to the subsequent PCR product purpose. Moreover, there are testing kits available with multiple enzymes that researchers can firstly use to test different performances on their samples (for example the PCR Enzyme Selection Kit from Thermofisher).

2.3.2. Low Template and Degraded Samples

The usefulness of mtDNA analysis in highly degraded samples is an advantage over other chromosomal profiling methods; still, dedicated amplification techniques have been implemented to maximize sensitivity in such circumstances. One of the most effective methods is the analysis of specific nucleotides, traditionally by SNaPshot Multiplex, where amplified fragments can be smaller than 100 bp. In fact, SNaPshot technique has been widely used since it is rapid, cost-effective, and can provide a haplogroup classification when control and coding region diagnostic positions are interrogated [19 - 22]. In some cases, the specific nucleotide screening may even provide sufficient information for a preliminary exclusion of samples involved in a forensic scenario [23] or be a simple and effective tool for diagnostics mitochondrial disorders. Nevertheless, the mtDNA complete amplification based on small fragments is costly, not feasible with traditional Sanger sequencing and is prone to co-amplification of NUMTs. Although generally in degraded samples nDNA is difficult to obtain, when compared to mtDNA, NUMT co-amplification must be considered. Specific methodologies such as SNaPshot may not allow the distinction between the two types of sequences.

Only recently with the development of MPS technologies it has been possible to successfully implement amplification strategies based on small amplicons, opening the door for the analysis of samples where in the past only a partial or no mitochondrial profile would be obtained [24, 25]. Several recent studies have applied these methodologies to sequence low quantity and low quality/degraded DNA from old or ancient samples (see following section) and recently, shotgun mtDNA sequencing protocols that can be routinely applied to forensic casework have also been published. Parson and collaborators (2015) developed a method for the analysis of mitogenomes from single hair shafts and/or shed hairs lacking follicular tissue. This methodology includes the amplification of 62 fragments ranging from 300 to 500 bp long and was tested through Illumina MiSeq instrument. Chaitanya and co-workers (2015) implemented a procedure to sequence the complete mitochondrial genome using short overlapping amplicons suitable for degraded DNA. In this study, 161 overlapping fragments were amplified, with sizes from 144-230 bp, and the procedure was validated in Ion Torrent Personal Genome Machine (PGM).

2.3.3. Ancient mtDNA Studies

The use of MPS-based sequencing contributed for an incredible rise of the available ancient mtDNA complete sequences. The use of traditional methods is expensive and time consuming; besides, ancient samples are also normally extremely rare and irreplaceable, and the use of a traditional sequencing approach requires a significant amount of sample material that cannot be reused. In contrast, by using MPS systems, the extracted DNA can be converted in libraries that can be used in later analysis [26]. Moreover, other technical approaches have revolutionized aDNA amplification, such as DNA hybridization enrichment [9, 27], being used to enrich MPS libraries for a specific target sequence before sequencing.

Furthermore, one of the most debated issues in using ancient hominin mtDNA is the problem of human contamination (exogenous DNA) and the ability to authenticate the ancient sequences reported. Contamination sources can include microbial DNA or other aDNA, however, in general, the major source of contamination is modern human DNA during sampling handling prior sequencing.

Apart from the good pre-laboratorial and laboratorial practices continuously encouraged, different methods to recognized contamination and distinct between endogenous and exogenous contributors have been explored [28]. These include, for instance, the analysis of postmortem DNA cytosine deamination damage (C to U/T changes at regular and 5'- methylated cytosines, respectively) [29 - 32]. Deamination mostly occurs at the ends of aDNA sequences and its frequency increases with the sample age. A study from Skoglund and collaborators (2014) has applied the above mentioned approach to successfully retrieve one complete mitogenome from a highly contaminated Neanderthal sample of Siberia (*Okladnikov* cave). In the same year a study from Meyer and collaborators (2014) also used cytosine deamination patterns to reconstruct an almost complete mitogenome from a hominin fossil with an estimate age of 400,000 years old in Atapuerca (*Sima de los Huesos* cave), Spain.

Apart from the above studies, at the moment several hominin mitochondrial aDNA sequences are available [33 - 37] and the continuous progress of MPS approaches with higher sensitivity is paving the way for future accomplishments in ancient DNA research.

2.4. MPS PLATFORMS

Commercially available MPS systems share common features but use different sequencing biochemistry approaches. Library preparation is the first required step for these high-throughput techniques. It includes the random fragmentation of DNA followed by the ligation of universal synthetic sequences (adapters), at both fragments ends. The adapter-ligated fragments obtained are *in situ* clonal amplified before sequencing on a solid surface (bead or glass microfluidic channel) with covalently attached adapter sequences that hybridize with the library adapters. Each platform has specific amplification and sequencing methods.

There are four main MPS sequencing biochemistry approaches: i) Pyrosequencing; ii) Sequencing by Synthesis; iii) Sequencing by Ligation; and iv) Ion Semiconductor Sequencing. The Single Molecule Sequencing Technology is also now commonly used (PacBio, Helicos and Nanopore systems), in which case a prior DNA fragment amplification it is not required, and in some cases the signal is also recorded in real-time (known as third generation sequencers).

Generally, after amplification, sequencing is performed by cyclic nucleotide incorporation steps with simultaneous detection of the base incorporated that can be for example through the detection of a fluorescent signal. The final stage comprises the analysis and alignment of the data collected during the sequencing process. Thousands to millions of parallel reactions occur during each instrument run and therefore enormous data sets are produced. For this reason, when compared to Sanger Sequencing, MPS platforms require a longer run time. Nevertheless, the yield of sequence reads and total bases per instrument run is significantly higher in MPS. The use of multiplexing also allows the simultaneous sequencing of a large number of samples during a single run (barcoding). For that, different libraries are prepared for each sample that is recognized by a demultiplexing algorithm in the final data analysis process.

Currently several MPS platforms are available (Table **1**), each with their own sequencing biochemistry and therefore their own advantages and limitations. Apart from those examples presented in Table **1**, other systems can be found in the market, including the Irys (BioNano Genomics), the Revolocity (Complete Genomics), the Helicos (SeqLL) or the Nanopore platform (Oxford Nanopore Technologies). A careful analysis is needed before selecting the suitable platform for a particular investigation. Each company usually provides different equipment, to fit different laboratory needs (different objectives and scales). In this context, the particular characteristics of the mitochondrial genome have to be taken into consideration. Specifically, mtDNA holds a high number of homopolymeric stretches and heteroplasmatic positions, together with a high GC content and dinucleotide repeats [38]. These chemical properties and polymorphic spots are important to understand the accuracy and the limits of detection of each MPS platform.

Compact MPS systems such as the Illumina MiSeq and Ion Torrent PGM, are becoming popular in research and diagnostic field. These two platforms are small in size and although having limited data throughput and higher per-base-cost than other systems such as Illumina HiSeq or Ion Proton, they are faster, versatile and

Table 1. Some of the actual available MPS platforms and their characteristics. The table was based on the information available in each manufacturer's webpages.

Manufacturer	Sequencing Method	Amplification Method	Platforms	Maximum Read Length (bp)*	Sequence Yield Throughput*	Run Time*	Reported Accuracy
Roche	Pyrosequencing	Emulsion PCR – Amplification on beads	**GS Junior**	400	35Mb	10 hours	99% accuracy at 400 bases and higher for preceding bases (Q20)
			GS Junior +	700	70Mb	18 hours	99% accuracy at 700 bases and higher for preceding bases (Q20)
			GS FLX Titanium XLR70	Up to 600	450 Mb	10 hours	99.995% consensus accuracy at 15x coverage
			GS FLX Titanium XL +	Up to 1.000	700 Mb	23 hours	99.997% consensus accuracy at 15x coverage
Pacific Biosciences	Sequencing by Synthesis	No clonal amplification Single molecule, real-time (SMRT) technology	**PacBio RS II and Sequel system**	Up to 60.000	Up to 1 Gb (per cell***)	30 minutes – 6 hours	> 99.999% consensus accuracy (Q50)

(Table 1) contd.....

Manufacturer	Sequencing Method	Amplification Method	Platforms	Maximum Read Length (bp)*	Sequence Yield Throughput*	Run Time*	Reported Accuracy
Illumina	Sequencing by Synthesis	Bridge PCR – Amplification on solid surface	**MiSeq**	2 x 300	Up to 15 Gb	4 - 56 hours	> 90% bases higher than Q30 at 1 x 36 bp > 70% bases higher than Q30 at 2 x 300 bp**
			NextSeq	2 x 150	Up to 120 Gb	11 - 29 hours	> 80% bases higher than Q30 at 1 x 75 bp > 75% bases higher than Q30 at 2 x 150 bp**
			HiSeq 2500	2 x 250	Up to 1 Tb	7 hours - 6 days	> 85% of bases above Q30 at 2×50 bp > 75% of bases above Q30 at 2×150 bp**
			HiSeq 3000	2 x 150	Up to 750 Gb	< 1 – 3.5 days	≥ 75% of bases above Q30 at 2 x 150 bp
			HiSeq 4000	2 x 150	Up to 1.5 Tb	< 1 – 3.5 days	≥ 75% of bases above Q30 at 2 x 150 bp
			HiSeq X Series (Ten and Five Systems)	2 x 150	Up to 1.8 Tb	< 3 days	≥ 75% of bases above Q30 at 2 x 150 bp

(Table 1) contd.....

Manufacturer	Sequencing Method	Amplification Method	Platforms	Maximum Read Length (bp)*	Sequence Yield Throughput*	Run Time*	Reported Accuracy
ThermoFisher Scientific (Life Technologies brand)	Semiconductor Sequencing	Emulsion PCR – Amplification on beads	**Ion PGM**	Up to 400	Up to 2 Gb	2.3 – 7.3 hours	> 99% aligned/measured accuracy
			Ion PGM Dx	200	Up to 1 Gb	4.4 hours	> 99% aligned/measured accuracy
			Ion Proton	Up to 200	Up to 10 Gb	2 – 4 hours	-
			Ion S5	Up to 400	Up to 15 Gb	2.5 – 4 hours	> 99% aligned/measured accuracy
			Ion S5 XL	Up to 400	Up to 15 Gb	2.5 – 4 hours	> 99% aligned/measured accuracy
ThermoFisher Scientific (Applied Biosystems brand)	Sequencing by Ligation	Emulsion PCR – Amplification on beads Emulsion PCR – Non-bead amplification (direct amplification on the FlowChip)	**5500**	75 bp (fragment)	Up to 15 Gb (per day)	24 hours	99.99% system accuracy
			5500xl	75 bp (fragment)	Up to 45 Gb (per day)	24 hours	99.99% system accuracy
			5500xl-W	1 x 50 bp (fragment)	Up to 120 Gb (per run)	-	Up to 99.999% system accuracy

*All the values are estimates; **Only the maximum and minimum manufacture available accuracy are herein reported; *From 1 to 16 SMRT Cells per run flexible, ideal for labs that do not require handling extremely large datasets [18, 39]. In fact, so far, the majority of the complete mtDNA available literature reports have used such platforms. Many of them focused in testing the sensitivity and coverage of such methodologies for specific situations. For these reasons, we pay special attention to these two particular platforms.

2.4.1. Illumina MiSeq™

The Illumina technology is based on Sequencing by Synthesis approach, where four fluorescent labeled nucleotides are incorporated by a DNA polymerase enzyme (Fig. **2**):

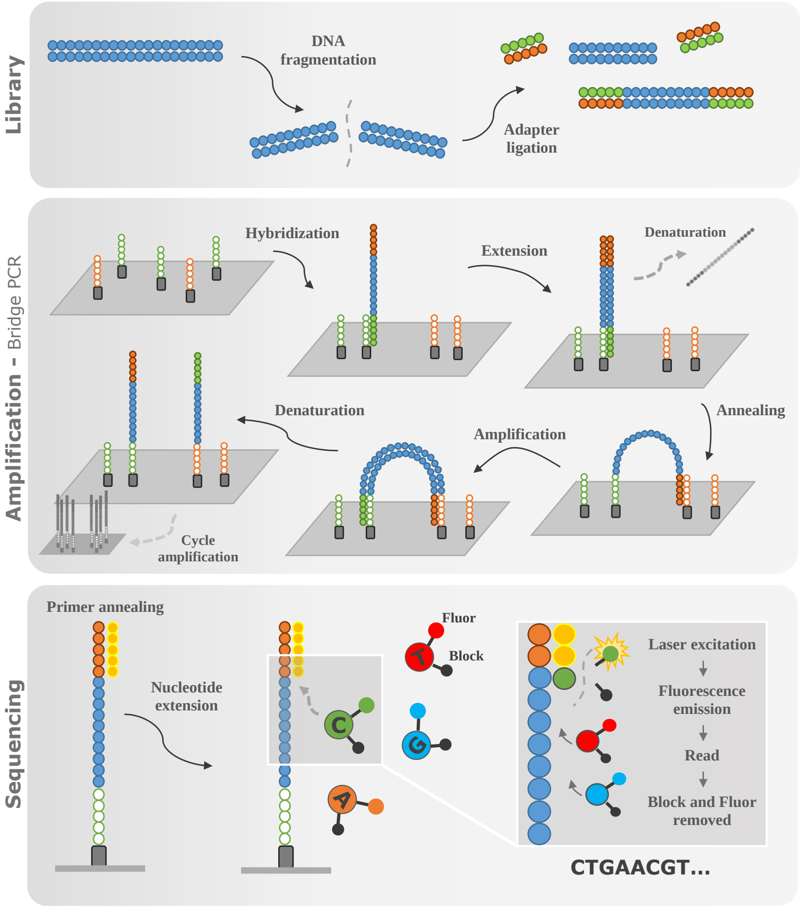


Fig. (2). Illumina sequencing and detection technology.

1. **Library preparation** - Template DNA is randomly fragmented followed by 5' and 3' adapters' ligation.
2. **Clonal amplification** - The fragments are immobilized in a flow cell. The surface of the cell is covered with complementary adapter sequences that allow DNA hybridization. A first stage of extension creates complementary template fragments linked to adapter-surface cluster. The following process is known as bridge amplification because each single-stranded immobilized fragment creates a bridge structure after the hybridization of the free end to other complementary adapter-surface cluster. Several cycles of bridge extension and denaturation creates random clusters of single-stranded DNA fragments on the surface.
3. **Sequencing and detection** - Sequencing starts with the annealing of a sequencing primer to the unbounded DNA end and nucleotide extension by a DNA polymerase. Each terminator-bound deoxynucleotide triphosphates (dNTPs) is labeled with a different fluorescent dye and a removable block. After incorporation fluorescence is recorded by a charge-coupled detector (CCD). The dye along with the terminal 3' blocker is chemically removed from DNA before each next cycle to allow incorporation of the succeeding dNTPs.
4. **Data analysis** - At the end all the sequence reads are aligned to a reference genome and interpreted.

2.4.2. Ion Torrent Personal Genome Machine (PGM™)

The PGM uses a Semiconductor system, a non-optical sequencing approach that detects slight changes in pH due to hydrogen ion release during nucleotide incorporation (Fig. **3**):

1. **Library preparation** - Template DNA is randomly fragmented followed by 5' and 3' adapters' ligation.
2. **Clonal amplification** - Library fragments are mixed with micron-scale beads covered by adapter complementary sequences. On the solution all the reagents needed for amplification and oil micelles are also present. This amplification strategy is known as emulsion PCR, since the mixture is shaken to form an emulsion where the amplicons and the beads are encapsulated in oil micelles (water-in-oil emulsion). Generally, due to low template concentrations each

micelle contains only one DNA library strand (or any) that hybridize with the bead surface. A first extension creates complementary template fragments linked to the adapter-surface bead. After several amplification cycles the bead is covered by template DNA and the oil is lastly separated from beads and the water solution (emulsion breaking).

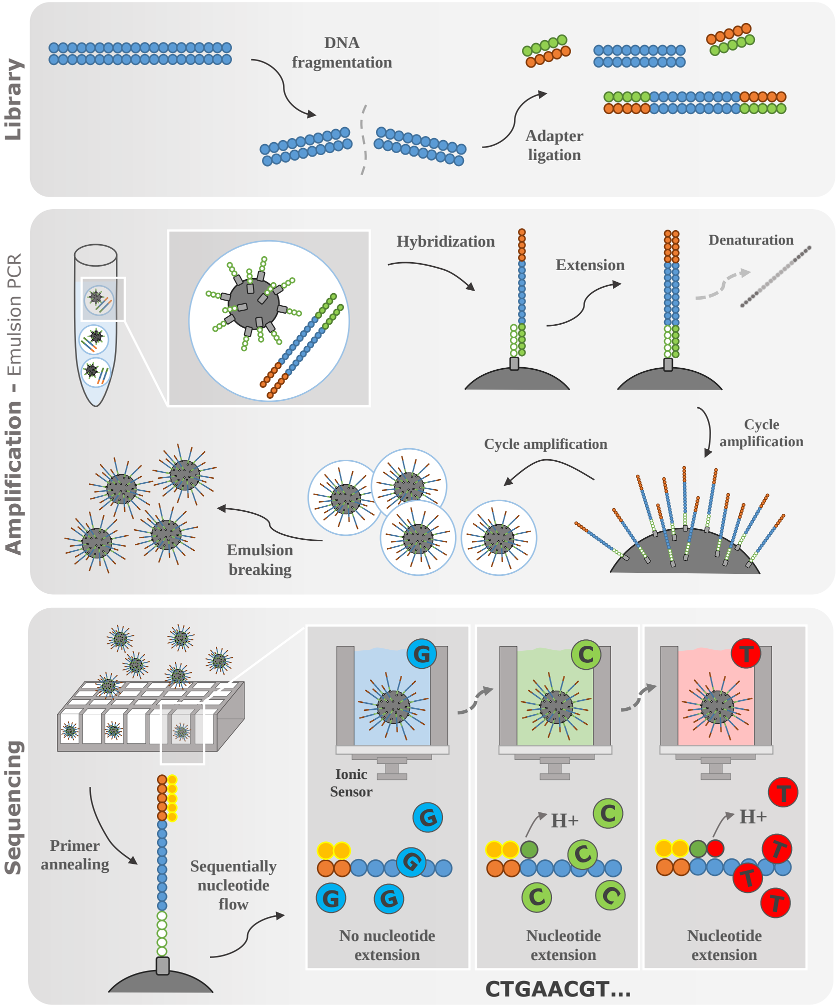


Fig. (3). Ion torrent sequencing and detection technology.

3. **Sequencing and detection** - The beads are loaded into microwells of an ion silicon chip and the DNA templates are primed for sequencing. Ideally only one amplicon-bead complex will be contained in each well. The sequencing begins when the well is flooded with dNTPs and DNA polymerases. Distinction between different bases is achieved by flows different nucleotides sequentially after another. When the corresponding nucleotide is inserted a single hydrogen ion (H^+) is released, generating a slight decrease of the pH that is immediately measured by the ionic sensor. When two equal nucleotides are added, the voltage will double, and the software records two identical bases.
4. **Data analysis** - Several sequencing reactions per chip are promoted. At the end the obtained data is grouped, filtered and the noise removed, ending with sequence report.

2.4.3. Platforms Comparisons

There are two main differences between Illumina and Ion Torrent platforms namely the amplification methodology employed (bridge and emulsion PCR, respectively), and the sequencing approach (Sequencing by Synthesis and Semiconductor Sequencing, respectively). Both have their own limitations in what concerns accuracy, read length, consumables, technical skills and informatics requirements, which will determine which strategy represents the best option for any given project. Globally, the major limitation of any MPS system besides the error rate is the Signal-to-Noise Ratio (SNR) during sequencing. Depending on the sequencing approach different noise contributors and errors are also commonly produced.

PGM was the first commercial sequencing machine to not require fluorescence and camera scanning, resulting in higher speed and lower cost. Contrariwise, the MiSeq requires a higher sequencing time but also has a higher throughput and a less labour intensive methodology [40]. It also integrates all the technology needed in a single compact equipment, including the data analysis. The major source of noise and sequencing errors (mainly substitutions errors) in Illumina technology is due to the use of a florescence-based sequencing approach. Incomplete blocking, which allow subsequent nucleotide incorporations, or the presence of residual fluorescence due to incomplete fluorescent dye cleavage can be some of the problems responsible for noise production [41]. The quality of the sequences in the Illumina technology also decreases with the increasing of the read length, due to consequent decay of florescence signal. In the PGM the use of native nucleotides minimizes the former problems and produces a more stable sequencing quality. The substitution errors occur at a lower frequency and the performance in sequences with high GC content is higher [39]. However, Illumina chemistry has a more accurate sequencing through the repetitive and/or homopolymeric regions, present in the mitochondrial genome. In fact, the error model of the PGM sequencing is defined largely by insertion or deletion errors that are also prevalent at the homopolymeric regions. The effect is more pronounced as the length of the homopolymer increases. To record the length of these regions, the sensor must detect the magnitude of the pH change to determine how many nucleotides were incorporated. An error occurs when the sensor erroneously quantifies the magnitude of the signal. Since in Illumina only a single base is incorporated in each cycle, its sequencing error rate in homopolymeric regions is lower [42].

Another important aspect in mtDNA analysis is the ability to precisely detect heteroplasmic positions, *i.e.* positions presenting a mixture of two alleles. Neutral variants are sometimes present in a heteroplasmic state, but many mitochondrial pathogenic mutations are heteroplasmic where the percentage of the mutated variant or allele play a role on the disease phenotype. Therefore adequate heteroplasmy detection and quantification is crucial. The traditional Sanger sequencing method is not sensitive enough to detect low-level heteroplasmy and, *per se*, does not allow for accurate allele quantification. It has been demonstrated that MPS technologies are more sensitive and more quantitative, particularly when contamination is controlled and the amplification strategy is designed carefully [43 - 45]. In fact, even using MPS, the main problem in heteroplasmy detection remains the distinction between real and artefactual heteroplasmic positions that can be caused by other samples contamination, NUMTs presence, or sequencing errors [46]. A study from Vancampenhout and co-workers (2014) demonstrated that the Miseq system might be more appropriate for low copy number heteroplasmy. The detection threshold for base variants identification was set on 2% for Miseq [45] and 5% for PGM machine [44]. The previous results on

PGM are also confirmed by Magalhães and collaborators (2015), where PGM has also been considered sensitive and accurate to detect mixture/heteroplasmy portions superiors to 5% in artificial heteroplasmic samples.

The ongoing evolution in MPS techniques on one hand increases the performance of the available platforms and on the other hand develops new systems. Understanding the limitations and advantages of each MPS platform is therefore not only useful to select the most appropriated method for each particular scientific experience, but also to understand the direction that needs to be undertaken for the improvement of the technology.

2.5. IMPLEMENTATION OF MPS IN THE ROUTINE RESEARCH AND DIAGNOSTIC LABORATORIES

Although it is becoming clear that MPS technology has advantages both in research and clinical laboratories, there are associated complexities to a successful implementation of these systems in the laboratory routine. A careful determination of which general strategy represents the best option for any particular application is still of utmost importance, either when using MPS or Sanger. The advantages of MPS systems are more conspicuous on large scale studies, where they allow to simplify and accelerate laboratory routine. Nonetheless, conventional sequencing methodologies such as Sanger will continue to be the strategy of choice for small scale studies in a near future due to its greater versatility [47]. Aspects such as the final sequencing costs, accuracy, reproducibility, quality and quantity of the samples and the data generated have also to be taken into consideration.

MPS systems are largely kit-based, thus increasing the final cost. It is possible to use alternatives not dependent of kits in some steps, preferentially in the more expensive ones, however these options can also decrease the accuracy of the data generated and the longer time spent in preparing reagents and protocol validation may not be compensatory from an economic point of view. Different platforms not only have different characteristics, but also different equipment and reagent costs that need to be considered. The decrease in sequencing cost is only significant if a careful strategy is applied, *i.e.*, when a compromise between high target length and number of samples analyzed exists. This is particularly important in mitochondrial studies, where the small low genome size (even for complete sequencing) requires the simultaneous analysis of several barcoded samples to permit reduction in the final sequencing cost. The need for a high computer capacity and storage to deal with the significant amount of data generated in MPS systems is also sometimes a drawback for laboratories, especially for small laboratories with restricted funds. All these issues must be prioritized before the acquisition and implementation of any MPS platform.

Validation procedures prior to the routine based-MPS are essential to ensure accuracy and high quality results. The implementation of strict laboratory guidelines and good practices are the first stage, which may include the accomplishment of any accreditation requirements and regular presence of professional guidance [48]. During validation the laboratories may need to test some parameters that include accuracy, sensitivity, specificity, precision and reproducibility. In clinical laboratories each test should also be validated and further parameters should be determined such as the depth of coverage or the ability to efficiently detect homopolymeric and heteroplasmic regions [38, 48]. Moreover, obtaining a high quality result is dependent on both sample handling/preparation and equipment features, but also on the development of appropriated analysis pipelines that should be correctly validated prior to sample analysis. The standardization and regulation of MPS tests is difficult due to the high number of different platforms with their own characteristics, but also due to the high range of different test possibilities [49]. Nevertheless, although challenging, one of the priorities of the field should be the standardization of MPS tests, from laboratory testing to final bioinformatics analysis, to contradict the currently observed internal laboratory regulation, sometimes substantially differentiated between laboratories, together with a continuous investigation and personal training that strictly follow the fast MPS technological progress.

CONCLUDING REMARKS

As platforms and sequencing techniques are rapidly evolving and new applications will arise. Single molecule sequencing technologies would possibly be the future direction of novel platforms. If a high accuracy can be accomplished and protocols are simplified this would decrease the possibility of errors during pre-amplifications steps. The problem of preferential amplification of some sequences over others is also prevented. Software tools for MPS analysis are also rapidly changing and adapting to the application requirements. Current and new products being developed cover different stages of the analysis including alignment of sequence reads, polymorphism detection and base-calling [47]. All these improvements will certainly encourage a broader use of such technologies, increasing their utilization in more areas and therefore, contributing for the enrichment of the scientific knowledge.

CONFLICT OF INTEREST

IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education. IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT, the Portuguese Foundation for Science and Technology. AG and LA are supported by FCT fellowship (grants SFRH/BPD/ 43646/2008 and SFRH/BPD/65000/2009, respectively), funded by POPH-QREN–Promotion of scientific employment, supported by the European Social Fund and national funds of the Ministry of Education and Science. SLM is supported by FCT (grant PTDC/ATP-DEM/4545/2012); financed by the European Social Funds (COMPETE-FEDER).

ACKNOWLEDGEMENTS

SLM, AG and LA drafted the chapter, edited for intellectual content and approved the final document.

REFERENCES

[1] Butler JM. Mitochondrial DNA analysis Forensic DNA typing: biology, technology, and genetics of STR markers. Elsevier Academic Press 2005; pp. 241-98.

[2] Hurst GDD, Jiggins FM. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc Biol Sci 2005; 272(1572): 1525-34.

[3] Hashiguchi K, Bohr VA, de Souza-Pinto NC. Oxidative stress and mitochondrial DNA repair: implications for NRTIs induced DNA damage. Mitochondrion 2004; 4(2-3): 215-22. [http://dx.doi.org/10.1016/j.mito.2004.05.014] [PMID: 16120387]

[4] Mollnau H, Wenzel P, Oelze M, *et al.* Mitochondrial oxidative stress and nitrate tolerance-

-comparison of nitroglycerin and pentaerythritol tetranitrate in Mn-SOD+/- mice. BMC Cardiovasc Disord 2006; 6(1): 44.
[http://dx.doi.org/10.1186/1471-2261-6-44] [PMID: 17092343]

[5] Parson W, Gusmão L, Hares DR, *et al.* DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet 2014; 13: 134-42.
[http://dx.doi.org/10.1016/j.fsigen.2014.07.010] [PMID: 25117402]

[6] Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet 2012; 13(12): 878-90.
[http://dx.doi.org/10.1038/nrg3275] [PMID: 23154810]

[7] Chinnery PF, Hudson G. Mitochondrial genetics. Br Med Bull 2013; 106(1): 135-59.
[http://dx.doi.org/10.1093/bmb/ldt017] [PMID: 23704099]

[8] Wong L-J. Next generation molecular diagnosis of mitochondrial disorders. Mitochondrion 2013; 13(4): 379-87.
[http://dx.doi.org/10.1016/j.mito.2013.02.001] [PMID: 23473862]

[9] Mamanova L, Coffey AJ, Scott CE, *et al.* Target-enrichment strategies for next-generation sequencing. Nat Methods 2010; 7(2): 111-8.
[http://dx.doi.org/10.1038/nmeth.1419] [PMID: 20111037]

[10] Parson W, Bandelt H-J. Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 2007; 1(1): 13-9.
[http://dx.doi.org/10.1016/j.fsigen.2006.11.003] [PMID: 19083723]

[11] Levin BC, Holland KA, Hancock DK, *et al.* Comparison of the complete mtDNA genome sequences of human cell lines HL-60 and GM10742A from individuals with pro-myelocytic leukemia and leber hereditary optic neuropathy, respectively, and the inclusion of HL-60 in the NIST human mitochondrial DNA standard reference material SRM 2392-I. Mitochondrion 2003; 2(6): 387-400.
[http://dx.doi.org/10.1016/S1567-7249(03)00010-2] [PMID: 16120335]

[12] Torroni A, Rengo C, Guida V, *et al.* Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 2001; 69(6): 1348-56.
[http://dx.doi.org/10.1086/324511] [PMID: 11595973]

[13] Fendt L, Zimmermann B, Daniaux M, Parson W. Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics 2009; 10(1): 139.
[http://dx.doi.org/10.1186/1471-2164-10-139] [PMID: 19331681]

[14] Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 2007; 24(3): 757-68.
[http://dx.doi.org/10.1093/molbev/msl209] [PMID: 17194802]

[15] Cui H, Li F, Chen D, *et al.* Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. Genet Med 2013; 15(5): 388-94.
[http://dx.doi.org/10.1038/gim.2012.144] [PMID: 23288206]

[16] Magalhães S, Marques SL, Alves C, *et al.* Evaluation of heteroplasmy detection in the Ion Torrent

PGM. Forensic Sci Int Genet 2015; 5: 13-5.
[http://dx.doi.org/10.1016/j.fsigss.2015.09.006]

[17] Ramos A, Santos C, Barbena E, *et al.* Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the new human genome reference sequence (GRCh37). Electrophoresis 2011; 32(6-7): 782-3.
[http://dx.doi.org/10.1002/elps.201000583] [PMID: 21425173]

[18] Jia H, Guo Y, Zhao W, Wang K. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. Sci Rep 2014; 4: 5737.
[http://dx.doi.org/10.1038/srep05737] [PMID: 25034901]

[19] Álvarez-Iglesias V, Jaime JC, Carracedo A, Salas A. Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 2007; 1(1): 44-55.
[http://dx.doi.org/10.1016/j.fsigen.2006.09.001] [PMID: 19083727]

[20] Ballantyne KN, van Oven M, Ralf A, *et al.* MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. Forensic Sci Int Genet 2012; 6(4): 425-36.
[http://dx.doi.org/10.1016/j.fsigen.2011.08.010] [PMID: 21940232]

[21] Coble MD, Just RS, O'Callaghan JE, *et al.* Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 2004; 118(3): 137-46.
[http://dx.doi.org/10.1007/s00414-004-0427-6] [PMID: 14760490]

[22] van Oven M, Vermeulen M, Kayser M. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. Investig Genet 2011; 2(1): 6.
[http://dx.doi.org/10.1186/2041-2223-2-6] [PMID: 21429198]

[23] Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo A, Parson W. Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 2006; 27(13): 2541-50.
[http://dx.doi.org/10.1002/elps.200500772] [PMID: 16721903]

[24] Parson W, Huber G, Moreno L, *et al.* Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Sci Int Genet 2015; 15: 8-15.
[http://dx.doi.org/10.1016/j.fsigen.2014.11.009] [PMID: 25438934]

[25] Chaitanya L, Ralf A, van Oven M, *et al.* Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA using the ion torrent personal genome machine. Hum Mutat 2015; 36(12): 1236-47.
[http://dx.doi.org/10.1002/humu.22905] [PMID: 26387877]

[26] Paijmans JL, Gilbert MT, Hofreiter M. Mitogenomic analyses from ancient DNA. Mol Phylogenet Evol 2013; 69(2): 404-16.
[http://dx.doi.org/10.1016/j.ympev.2012.06.002] [PMID: 22705825]

[27] Summerer D. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics 2009; 94(6): 363-8.
[http://dx.doi.org/10.1016/j.ygeno.2009.08.012] [PMID: 19720138]

[28] Knapp M, Lalueza-Fox C, Hofreiter M. Re-inventing ancient human DNA. Investig Genet 2015; 6(1): 4.

[http://dx.doi.org/10.1186/s13323-015-0020-4] [PMID: 25937886]

[29] Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 2013; 29(13): 1682-4. [http://dx.doi.org/10.1093/bioinformatics/btt193] [PMID: 23613487]

[30] Meyer M, Fu Q, Aximu-Petri A, *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 2014; 505(7483): 403-6. [http://dx.doi.org/10.1038/nature12788] [PMID: 24305051]

[31] Skoglund P, Malmström H, Raghavan M, *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 2012; 336(6080): 466-9. [http://dx.doi.org/10.1126/science.1216304] [PMID: 22539720]

[32] Skoglund P, Northoff BH, Shunkov MV, *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci USA 2014; 111(6): 2229-34. [http://dx.doi.org/10.1073/pnas.1318934111] [PMID: 24469802]

[33] Briggs AW, Good JM, Green RE, *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009; 325(5938): 318-21. [http://dx.doi.org/10.1126/science.1174462] [PMID: 19608918]

[34] Dabney J, Knapp M, Glocke I, *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA 2013; 110(39): 15758-63. [http://dx.doi.org/10.1073/pnas.1314445110] [PMID: 24019490]

[35] Green RE, Malaspinas A-S, Krause J, *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 2008; 134(3): 416-26. [http://dx.doi.org/10.1016/j.cell.2008.06.021] [PMID: 18692465]

[36] Krause J, Briggs AW, Kircher M, *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol 2010; 20(3): 231-6. [http://dx.doi.org/10.1016/j.cub.2009.11.068] [PMID: 20045327]

[37] Krause J, Fu Q, Good JM, *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 2010; 464(7290): 894-7. [http://dx.doi.org/10.1038/nature08976] [PMID: 20336068]

[38] Wong L-J. Next-generation sequencing analyses of the whole mitochondrial genome. In: Wong L-J, Ed. Next Generation Sequencing. New York: Springer 2013; pp. 203-19. [http://dx.doi.org/10.1007/978-1-4614-7001-4_11]

[39] Liu L, Li Y, Li S, *et al.* Comparison of next-generation sequencing systems. Biomed Res Int 2012. [http://dx.doi.org/10.1155/2012/251364]

[40] Seo S, Zeng X, Assidi M, *et al.* High throughput whole mitochondrial genome sequencing by two platforms of massively parallel sequencing. BMC Genomics 2014; 15 (Suppl. 2): 7. [http://dx.doi.org/10.1186/1471-2164-15-S2-P7] [PMID: 24384011]

[41] Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto, Calif) 2013; 6: 287-303. [http://dx.doi.org/10.1146/annurev-anchem-062012-092628] [PMID: 23560931]

[42] Mascher M, Wu S, Amand PS, Stein N, Poland J. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. PLoS One 2013; 8(10): e76925.
[http://dx.doi.org/10.1371/journal.pone.0076925] [PMID: 24098570]

[43] Parson W, Strobl C, Huber G, *et al.* Reprint of: Evaluation of next generation mtGenome sequencing using the ion torrent personal genome machine (PGM). Forensic Sci Int Genet 2013; 7(6): 632-9.
[http://dx.doi.org/10.1016/j.fsigen.2013.09.007] [PMID: 24119954]

[44] Seneca S, Vancampenhout K, Van Coster R, *et al.* Analysis of the whole mitochondrial genome: translation of the Ion Torrent Personal Genome Machine system to the diagnostic bench? Eur J Hum Genet 2015; 23(1): 41-8. [quest].
[http://dx.doi.org/10.1038/ejhg.2014.49] [PMID: 24667782]

[45] Vancampenhout K, Caljon B, Spits C, *et al.* A bumpy ride on the diagnostic bench of massive parallel sequencing, the case of the mitochondrial genome 2014; 9(10): e12950.
[http://dx.doi.org/10.1371/journal.pone.0112950]

[46] Just RS, Irwin JA, Parson W. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int Genet 2015; 18: 131-9.
[http://dx.doi.org/10.1016/j.fsigen.2015.05.003] [PMID: 26009256]

[47] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26(10): 1135-45.
[http://dx.doi.org/10.1038/nbt1486] [PMID: 18846087]

[48] Lubin I, Kalman L, Gargis A. Guidelines and approaches to compliance with regulatory and clinical standards: Quality control procedures and quality assurance. In: Wong L-J, Ed. Next Generation Sequencing. New York: Springer 2013; pp. 255-73.
[http://dx.doi.org/10.1007/978-1-4614-7001-4_14]

[49] Veltman JA, Cuppen E, Vrijenhoek T. Challenges for implementing next-generation sequencing-based genome diagnostics: it's also the people, not just the machines. Per Med 2013; 10(5): 473-84.
[http://dx.doi.org/10.2217/pme.13.41]

CHAPTER 3

Somatic *vs* Germinal Mutations in Mitochondrial DNA: Is There Any Relation with Human Health and Aging?

Amanda Ramos[1,2,3,4,*], **Maria Pilar Aluja**[1], **Manuela Lima**[2,3,4] and **Cristina Santos**[1,*]

[1] *Unitat Antropologia Biològica, Department BABVE, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain*

[2] *Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal*

[3] *Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal*

[4] *"Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal*

Abstract: Mitochondrial DNA (mtDNA) heteroplasmy is an almost universal condition in humans. The proportion of heteroplasmic mtDNA mutations that is heritable rather than accumulated during life has, however, remained almost unknown. The main goal of this work was to investigate the contribution of germinal *versus* somatic heteroplasmy, exploring its impact on health and aging. Blood samples from 101 individuals were previously used to generate full mtDNA sequences. Taking into account the embryonic origin of the tissues and the heterogeneity of site specific mutation rate of mtDNA robust criteria of heteroplasmy classification was applied. The mtDNA regions encompassing the 28 heteroplasmic positions detected in blood samples were sequenced in buccal epithelial samples as a reference from an alternative tissue with different embryonic origin. Based on the proposed classification data published by Li *et al.* (2015) was reanalyzed. Moreover, the predicted functional impact of non-synonymous mutations was evaluated. Most of heteroplasmies detected were germinal or somatic prior gastrulation and most of the somatic heteroplasmies were present in a single tissue. Somatic heteroplasmies were mostly present in older individuals, suggesting that they could be related to aging process. Three out of five

[*] **Corresponding authors Amanda Ramos and Cristina Santos:** Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain; Tel/Fax: +34935811503/+34935811321; E-mail: amanda.ramos.reche@gmail.com, cristina.santos@uab.cat

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

non-synonymous mutations in heteroplasmy (all of them classified as germinal or somatic prior gastrulation) occurred in highly conserved positions, presenting a probability >60% of being deleterious. Although germinal heteroplasmies (or somatic prior gastrulation) can contribute to the development of disease and to the aging process, most of the heteroplasmies detected in both studies present a level of the alternative allele frequency below 60%, likely not affecting fitness and escaping selection.

Keywords: Age, Disposable soma theory, Embryonic origin, Germinal, Health, Heteroplasmy, MtDNA, Mutation accumulation theory, Purifying selection, Somatic.

3.1. INTRODUCTION

Mitochondria play a critical role in the genesis of cellular energy. These organelles harbor several copies (between 100 and 10000, depending on the cell type) of a compact and independently replicating genome (mtDNA) [1]. The presence of several mtDNA copies within an individual can lead to an heterogeneous population of mtDNAs within the same cell, and even within the same mitochondrion, a condition known as heteroplasmy [1].

The study of human mitochondrial heteroplasmy goes back to the 1980s, and was first identified in genetic studies of mitochondrial diseases [2, 3]. However, heteroplasmy is also present in normal individuals. In fact, the emergence of new sequencing platforms has improved the resolution of mitochondrial genomes, offering the opportunity to recognize that mitochondrial heteroplasmy is an almost universal condition in humans [4]. Current estimations in the European population indicate that over 60% of individuals have more than one mtDNA type, the frequency of point heteroplasmy being of 28.7% [5, 6]. Recently, using data from the 1000 Genomes, Ye *et al.* [7] demonstrated that nearly 90% of the individuals carry at least one heteroplasmy. Heteroplasmy can be the result of novel germinal mutations that arise within a maternal lineage, but can also be the outcome of somatic mutations, which can arise at different moments of the individual development and during his entire life.

It has been generally accepted that a low level of heteroplasmy does not impact mitochondrial function; once the level of mutant mtDNA exceeds a certain threshold, however, the phenotypic consequences can become evident [8]. In this sense, recent evidence suggest that pathogenic mutations in heteroplasmy are prevalent in healthy individuals, although the levels of heteroplasmy are kept below the threshold [7, 9, 10]. Moreover, studies using mouse models [10] have shown that admixtures of two different mtDNAs, even if they are equally efficient, can be genetically unstable and associated with metabolic, behavioral and cognitive alterations. In the same line of evidence, heteroplasmic mice also developed systemic hypertension, had increased body and fat mass and displayed abnormalities concerning electrolytes and hematological parameters [9]. The results of Acton *et al.* [9] clearly indicate that although the heteroplasmic mice appeared to be healthy, there were several underlying abnormalities in standard physiological and metabolic parameters [9]. Ye *et al.* [7] raised up also the concern that mitochondrial heteroplasmy carried by healthy individuals could undergone clonal expansion, leading to high frequency of pathogenic variants later in life and, by consequence, to age-related diseases.

The presence of heteroplasmy has been commonly associated with aging and degenerative diseases, due to a decline in mitochondrial function occurring in both these processes (for a review see [7, 11 - 14]). On the other hand, it has been suggested that some heteroplasmic positions may increase longevity [15, 16]. In this context, the observation that mtDNA somatic mutations accumulate during the individual’s lifetime is one of the evidences that lead to the formulation of different mechanistic and evolutionary theories of aging [17]. The proposal that reactive oxygen species (ROS) drive aging by creating cumulative damage to vital cellular processes dates back to 1950s [18]. According to this proposal, mitochondria are responsible for the oxidative stress generated as a result of the Oxidative Phosphorylation (OXPHOS), leading a higher exposure of mtDNA to mutagenic events as compared to nuclear DNA (nDNA). The exposition to ROS would induce the generation of new somatic mutations in heteroplasmy, which would then increase their frequency with age. Notwithstanding, there are several lines of evidence to support the hypothesis that most mtDNA mutations are generated by replication errors and that oxidative damage is not a major contributor [19 - 21]. In fact, it has been postulated that ROS generation is not a cause of aging, but rather represents a stress signal in response to age-dependent damage [22]. Moreover, slipped mispairing during mtDNA replication has been suggested to be an important mechanism for formation of mtDNA deletions [23]; in accordance with such mechanisms, most human mtDNA deletions are indeed flanked by repeated sequences [24].

The proportion of heteroplasmic mtDNA mutations in the whole mitochondrial genome that is heritable rather than accumulated during life has remained almost unknown at the population level. However, this may be transcendent to understand the contribution of mtDNA mutations to disease and aging. In this sense, Li *et al.* [25] published the so far most comprehensive study of mtDNA heteroplasmy variation across different tissues. The authors classified the heteroplasmic point mutations as inherited or somatic, taking into account their distribution in 12 different tissues: a) heteroplasmies present in a single tissue were classified as somatic, b) heteroplasmies in three or more tissues as inherited and c) heteroplasmies present in two tissues as a mixture of inherited and somatic. According to Li *et al.* [25], around 43-65% of heteroplasmies would be inherited. Although the authors reported these frequencies as a rough guide, concerns about the criteria used to classify the origin of the heteroplasmic mutations can be raised, since the embryonic origin of tissues was not considered for mutation classification. Moreover, the authors [25] did not provide information about how each tissue was isolated and the hypothesis that the intestines and skin do not have traces of tissues of a mesodermic origin cannot be excluded.

In the present work the complete mitochondrial genome of one hundred and one individuals and the reanalysis of the data from Li *et al.* [25] have been performed. The main goals were: i) to investigate the contribution of germinal *versus* somatic mitochondrial heteroplasmy; and ii) to explore the impact of germinal/somatic heteroplasmy in health and aging.

3.2. MATERIAL AND METHODS

3.2.1. Sample Selection

One hundred and one blood and buccal epithelial samples from maternally unrelated Spanish individuals were used in the present study. Samples were anonymized in all steps of the laboratorial procedures. All information was managed under strictly confidential circumstances and participants signed a written consent form. The present study was approved by the Ethics Committee of the Autonomous University of Barcelona.

3.2.2. MtDNA Analysis

The whole mitochondrial genome from blood samples was previously sequenced and heteroplasmy detected and authenticated [5]. In brief, the mitochondrial genome was amplified and sequenced using a set of nine primer pairs that prevents nuclear insertions of mitochondrial origin (NUMTs) coamplification, as previously described [26, 27]. The sequences obtained by capillary electrophoresis were analyzed and aligned in relation to the revised Cambridge Reference Sequence (CRS) [28]. The authentication of mtDNA heteroplasmy was performed following three main steps: 1) DNA extraction, PCR amplification, and sequencing of mtDNA; 2) Sequences revealing heteroplasmy in step 1 were confirmed by a second amplification and sequencing reactions; 3) Samples presenting heteroplasmy in step 2 were amplified and sequenced using an independent DNA extract. Levels of heteroplasmy were determined as described in Santos *et al.* [29] using the average of height of peaks in the electropherograms. In accordance to the sensitivity of the sequencing technique previously evaluated by Ramos *et al.* [5] the heteroplasmy threshold was 10%.

Regions containing heteroplasmic positions previously detected and authenticated in blood samples were amplified and sequenced in buccal epithelial cells following the same strategy. Total DNA from buccal epithelial cells was extracted using JETQUICK Blood DNA Spin Kit (Genomed, Löhne, Germany) according to the manufacturer's specifications.

3.2.3. Classification of mtDNA Heteroplasmy in Somatic or Germinal

Taking into account the embryonic origin of the tissues and the heterogeneity of site specific mutation rate of mtDNA a robust criteria of heteroplasmy classification is proposed: 1) Germinal or somatic prior to gastrulation: heteroplasmy present in all embryonic layers (EL) or heteroplasmy present in two EL with low alternative allele frequency (AAF) (mean <0.08); 2) Early somatic after gastrulation: heteroplasmy present in two or more tissues within the same EL in a non-hotspot position (as previously defined by Ramos *et al.* [5]); 3) Recurrent somatic: heteroplasmy present in two tissues in a hotspot position or heteroplasmy located in a stable position that appears as hotspot for a specific tissue; and 4) Somatic: heteroplasmy present in one tissue. This classification assumes a strict isolation of tissues; a possible blood contamination in highly vascularized tissues, cannot, however, be excluded.

In order to establish the origin of point heteroplasmies detected in blood samples from the present study, buccal epithelial samples of the same individuals were used as a reference of another tissue with different embryonic origin (blood as a reference tissue of mesodermal origin and buccal cells of a tissue of ectodermic origin). Moreover, both blood and buccal epithelial cells present relatively high cell turnovers [30]. The heteroplasmic state of the point mutations was classified as germinal or somatic depending on the presence or not of heteroplasmy in both blood and buccal epithelial cells. Mutations were considered germinal or somatic prior gastrulation if heteroplasmy was present in the two tissues and were considered somatic if they were present only in blood.

Based on the proposed classification data published by Li *et al.* [25] was reevaluated; tissues were classified as endodermic (large intestine, liver and small intestine), mesodermic (blood, kidney, myocardial muscle and skeletal muscle) and ectodermic (cerebellum, cerebrum, cortex and skin).

3.2.4. *In Silico* Prediction of Functional Impact of Mutations and Data Analysis

To predict the functional impact of mutations of the present data, the number of hits in the mtDNA phylogeny, the frequency in the population database, and the conservation index (CI), both at nucleotide and at amino acid level, were previously calculated [5]. Moreover, the functional impact of non-synonymous mutations was further evaluated using the software package MutPred [31]. A stable position was defined when presented ≤ 1 hit in the mtDNA phylogeny, a low distribution in the population, high nucleotide and amino acid CI and high probability of deleteriousness.

The analyses of the relation between heteroplasmy and age were performed for those individuals whose information on age was available. Statistical comparison of age between germinal and somatic heteroplasmy was performed with the Mann-Whitney test using the program SPSS ver. 15.0.1 software [32].

Table 1. Complete results of heteroplasmic positions analyzed in the present study according to their origin.

Heteroplasmy Origin		**Germinal or Somatic Prior Gastrulation**	**Somatic**
Number of heteroplasmies (%)		24 (85.7)	4 (14.3)
Mean age (years)		58.7	69.2
Location Of Heteroplasmy	Protein coding genes - N (%)	13 (54.2)	2 (50.0)
	tRNAs and rRNAs - N (%)	5 (20.8)	0 (0)
	Control region - N (%)	6 (25.0)	2 (50.0)
	Coding:non coding ratio	03:01	01:01
Amino Acid Change	S - N (%)	9 (37.5)	1 (25.0)
	NS - N (%)	4 (16.7)	1 (25.0)
	NS-S ratio	1:2.25	01:01
Stability	≤1 hits phylogeny - N (%)	10 (41.6)	2 (50.0)
	Low distribution in population[a] - N (%)	19 (79.2)	3 (75)
	High Nucleotide CI[b] - N (%)	3 (12.5)	0 (0)
	High probability of deleteriousness[c] - N (%)	3 (12.5)	0 (0)
	All Stability Items - N (%)	3 (12.5)	0 (0)

S: synonymous, NS: non-synonymous; CI: conservation index; [a] Distribution of the minor variant in population database <5%; [b] CI >92%; [c] >60%

3.3. RESULTS

3.3.1. Frequency and Nature of Heteroplasmy

The analysis of the whole mtDNA sequences obtained from blood samples of the 101 individuals previously studied [5] revealed the presence of point heteroplasmy in 24 individuals. From these, 20 presented 1 heteroplasmic position, 3 presented 2 heteroplasmic positions and 1 presented 3 heteroplasmic positions. In total, 29 heteroplasmic positions were detected. For 23 of the heteroplasmic individuals a sample of buccal cells was also available, which allowed the sequencing of the mtDNA regions encompassing 28 heteroplasmic positions detected. From the 28 point heteroplasmies detected in blood samples, 24 were also present in buccal cells, whereas only 4 positions did not show any signal of mix variant (Table **1**). Thus, 85.7% of the point heteroplasmies detected in the present study are germinal and only 14.3% are of somatic origin (Table **1**).

The reevaluation of data from Li *et al.* [25] reveal that a 42.4% of heteroplasmies are classified as germinal or somatic prior gastrulation and most of the somatic heteroplasmies are present in a single tissue (Table **2**). Moreover, the mean AAF is significantly higher in heteroplasmies with a germinal origin or somatic prior gastrulation (mean= 0.076) when compared to somatic (mean= 0.043) (Mann-Whitney test *P*<0.001).

Table 2. Classification of heteroplasmies reported by Li *et al.* [25] considering the embryonic origin of tissues and the heterogeneity of site specific mutation rate. (EL: Embryonic layer, AAF: Alternative allele frequency).

	Classification of Heteroplamies	N	Frequency (%)	Mean of AAF
1.	**Germinal or somatic prior gastrulation** Heteroplasmy present in all EL Heteroplasmy present in 2 EL with low AAF (mean <0.08)	508#	42.4	0.076
2.	**Early somatic after gastrulation** Heteroplasmy present in 1 EL (≥2 tissues) in a non-hotspot position§	5	0.4	0.05
3.	**Recurrent somatic** Heteroplasmy present in 2 tissues in a hotspot position§	187	15.6	0.047
	3*. Recurrent somatic (hotspot specific-tissue) Heteroplasmy in a stable position§ that appears as hotspot for a specific tissue	77	6.4	0.028
4.	**Somatic** Heteroplasmy present in 1 tissue	421	35.1	0.044

§ As previously defined by Ramos *et al.* [5]

Including 10 heteroplasmies in which we cannot ruled out the possibility of blood contamination due to tissue vascularization

*Positions involved: 60 (N=37), 408 (N=19) and 564 (N=20)

To compare both studies, heteroplasmies detected by Li *et al.* [25] in blood with an AAF >8% (our limit of sensitivity) were analyzed (47 heteroplasmies), revealing a high frequency of heteroplasmies with a germinal origin or somatic prior gastrulation (95.7%). Thus showing that germinal heteroplasmy is the dominant one if heteroplasmies with an AAF >8% were considered.

3.3.2. Functional Impact of Heteroplasmy

An extensive analysis concerning the stability of all heteroplasmic positions detected by us has been previously performed [5]. In addition, in the present study, an *in silico* predictive analysis for all non-synonymous mutations was performed. From 20 point heteroplasmies located in the coding region, 18 were classified as germinal and 2 as somatic (Table **1**). From these, 5 point heteroplasmic positions were non-synonymous, 4 were germinal and 1 was somatic (Table **1**). Three out of these five heteroplasmic positions were found to be highly stable and conserved (6054, 7697 and 8603), having a nearly null representation of the minor variant in the mtDNA phylogeny or in the population database and presenting a nucleotide CI over 90% (Table **1**). Moreover, the *in silico* prediction of the functional impact of non-synonymous mutations revealed that all the highly stable and conserved positions presented a probability higher than 60% of being deleterious (Table **1**). Noteworthy, all of these stable heteroplasmic positions have a germinal origin whereas none of somatic heteroplasmies are located in stable positions.

Combining the present data with that from Li *et al.* [25], the average age for somatic heteroplasmy (70.4 years ± 16.06) was significantly higher than the one reported for germinal heteroplasmy (63.4 years ± 21.65) (Mann-Whitney test: $p<0.001$).

3.4. DISCUSSION

The majority of mutations in heteroplasmy detected in the present study were present in both blood and buccal cells, indicating that an important fraction of mutations detected in heteroplasmy are germinal or originated in very early stages of the development, before gastrulation and the formation of the three primary germinal layers (ectoderm, endoderm and mesoderm). This result is corroborated by the classification performed in the present study of nearly 1200 heteroplasmies published by Li *et al.* [25].

The elevated frequency of point heteroplasmy with germinal origin is in accordance with results previously reported by others [4, 29, 33]. Santos *et al.* [29], analyzed the control region of mtDNA in extended families, confirming the origin of point heteroplasmy. These previous authors reported a frequency of 61.6% families with germinal heteroplasmies. A similar result was reported by Payne *et al.* [4], which detected 4-fold more variants transmitted between first degree maternal relatives than mutations with somatic origin. Considering studies based on unrelated individuals that used tissues comparable to those used in the present study, Andrew *et al.* [33] detected a frequency of 80% of germinal mutations in the control region. Similar results were detected in the present study when only the control region is considered (75% of mutations were germinal).

The evidence that heteroplasmy is an almost universal condition, combined with the fact that in more than 80% of the cases the heteroplasmy is germinal, could have important implications, since individuals have a mixture of mtDNAs during their entire lives virtually in all their cells. As previously stated, studies in mouse models [10] have shown that admixtures of two different mtDNAs can be genetically unstable and have been associated with several alterations that may have possible clinical effects. Moreover, recent evidence suggest that maternally transmitted mtDNA mutations may play a role in aggregating aspects of normal human aging and pathology, contributing to a baseline mutation load that could induce pathogenicity once the threshold is reached [34]. It is possible that the presence of germinal heteroplasmy in humans would impair health of the individuals and ultimately contribute to aging and to the development of disease, and this point deserves future investigation. This idea is even more reinforced in front of the present data, since a high representation of point heteroplasmies with germinal origin are located in stable positions and with a relatively high level of pathogenicity.

It has long been proposed that germ cells are endowed with special maintenance and repair systems [35, 36]. In this sense, a bottleneck coupled with an effective quality screen might select embryos that carry only intact mitochondria. The persistence of germinal heteroplasmy during life evidenced from the present data and data from Li *et al.* [25] in tissues with high cell turnover, in which selection could potentially act eliminating heteroplasmy, would be in accordance to the

Disposable Soma Theory [35]; given finite resources, the more an animal expends on bodily maintenance, the less it can expend on reproduction, and *vice versa*. In this sense, the persistence of germinal heteroplasmy would save resources for reproduction even if it is harmful later in life contributing to the development of disease and to aging process. Thus, although germinal heteroplasmy can contribute to the development of disease and to the aging process, it is probable that germinal heteroplasmy has no effect on the fitness and is not selected against. The deleteriousness associated to a particular heteroplasmic position is determined by the level of the pathogenic variant and it has been recently proposed that 60% would be a good estimate of the threshold [7]. In this sense, most of the heteroplasmies detected in both studies present a level of the AAF below 60%. Thus, it seems that purifying selection would only act reducing levels of pathogenicity when derived allele frequency is larger than 60%.

It appears that somatic mutations could be related to aging/longevity since they are present in older individuals, in accordance with the Mutation Accumulation Theory [37]. It is often assumed that the age-associated somatic mtDNA mutations are generated due to damage and therefore start to accumulate in aging adults. However, an alternative possibility is that somatic mtDNA mutations are generated by replication errors and that many of the somatic mtDNA mutations in adults can be traced back to embryonic development or early postnatal life [38].

In summary, although germinal heteroplasmies (or somatic prior gastrulation) can contribute to the development of disease and to the aging process, most of the heteroplasmies detected in both studies present a level of the AAF below 60% and it is likely that they have no effect on fitness and are not selected against.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

This work was supported by MEC (projects: CGL2009-08205; CGL2006-07374; CGL2008-03408-E) and by Generalitat de Catalunya (Ref. 2009SGR566, 2014SGR1420). AR is a postdoctoral fellow of Fundo Regional para a Ciência (M3.1.7/F/031/2011).

REFERENCES

[1] Chinnery PF, Hudson G. Mitochondrial genetics. Br Med Bull 2013; 106: 135-59. [http://dx.doi.org/10.1093/bmb/ldt017] [PMID: 23704099]

[2] Holt IJ, Harding AE, Morgan-Hughes JA. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. Nature 1988; 331(6158): 717-9. [http://dx.doi.org/10.1038/331717a0] [PMID: 2830540]

[3] Wallace DC, Zheng XX, Lott MT, *et al.* Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease. Cell 1988; 55(4): 601-10. [http://dx.doi.org/10.1016/0092-8674(88)90218-8] [PMID: 3180221]

[4] Payne BA, Wilson IJ, Yu-Wai-Man P, *et al.* Universal heteroplasmy of human mitochondrial DNA. Hum Mol Genet 2013; 22(2): 384-90. [http://dx.doi.org/10.1093/hmg/dds435] [PMID: 23077218]

[5] Ramos A, Santos C, Mateiu L, *et al.* Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. PLoS One 2013; 8(10): e74636. [http://dx.doi.org/10.1371/journal.pone.0074636] [PMID: 24098342]

[6] Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 2010; 87(2): 237-49. [http://dx.doi.org/10.1016/j.ajhg.2010.07.014] [PMID: 20696290]

[7] Ye K, Lu J, Ma F, Keinan A, Gu Z. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. Proc Natl Acad Sci USA 2014; 111(29): 10654-9. [http://dx.doi.org/10.1073/pnas.1403521111] [PMID: 25002485]

[8] Rossignol R, Faustin B, Rocher C, Malgat M, Mazat JP, Letellier T. Mitochondrial threshold effects. Biochem J 2003; 370(Pt 3): 751-62. [http://dx.doi.org/10.1042/bj20021594] [PMID: 12467494]

[9] Acton BM, Lai I, Shang X, Jurisicova A, Casper RF. Neutral mitochondrial heteroplasmy alters physiological function in mice. Biol Reprod 2007; 77(3): 569-76. [http://dx.doi.org/10.1095/biolreprod.107.060806] [PMID: 17554081]

[10] Sharpley MS, Marciniak C, Eckel-Mahan K, *et al.* Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. Cell 2012; 151(2): 333-43. [http://dx.doi.org/10.1016/j.cell.2012.09.004] [PMID: 23063123]

[11] Lagouge M, Larsson NG. The role of mitochondrial DNA mutations and free radicals in disease and ageing. J Intern Med 2013; 273(6): 529-43. [http://dx.doi.org/10.1111/joim.12055] [PMID: 23432181]

[12] Wallace DC. Mitochondrial DNA mutations in disease and aging. Environ Mol Mutagen 2010; 51(5): 440-50.

[PMID: 20544884]

[13] Kennedy SR, Loeb LA, Herr AJ. Somatic mutations in aging, cancer and neurodegeneration. Mech Ageing Dev 2012; 133(4): 118-26.
[http://dx.doi.org/10.1016/j.mad.2011.10.009] [PMID: 22079405]

[14] Kazachkova N, Raposo M, Montiel R, *et al.* Patterns of mitochondrial DNA damage in blood and brain tissues of a transgenic mouse model of Machado-Joseph disease. Neurodegener Dis 2013; 11(4): 206-14.
[http://dx.doi.org/10.1159/000339207] [PMID: 22832131]

[15] Rose G, Passarino G, Scornaienchi V, *et al.* The mitochondrial DNA control region shows genetically correlated levels of heteroplasmy in leukocytes of centenarians and their offspring. BMC Genomics 2007; 8: 293.
[http://dx.doi.org/10.1186/1471-2164-8-293] [PMID: 17727699]

[16] Salvioli S, Capri M, Santoro A, *et al.* The impact of mitochondrial DNA on human lifespan: a view from studies on centenarians. Biotechnol J 2008; 3(6): 740-9.
[http://dx.doi.org/10.1002/biot.200800046] [PMID: 18548739]

[17] Goldsmith TC, Ed. The evolution of aging - how new theories will change the future of medicine. Crownsville, MD: Azinet Press 2013.

[18] Harman D. Aging: a theory based on free radical and radiation chemistry. J Gerontol 1956; 11(3): 298-300.
[http://dx.doi.org/10.1093/geronj/11.3.298] [PMID: 13332224]

[19] Zheng W, Khrapko K, Coller HA, Thilly WG, Copeland WC. Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. Mutat Res 2006; 599(1-2): 11-20.
[http://dx.doi.org/10.1016/j.mrfmmm.2005.12.012] [PMID: 16490220]

[20] Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet 2013; 9(9): e1003794.
[http://dx.doi.org/10.1371/journal.pgen.1003794] [PMID: 24086148]

[21] Itsara LS, Kennedy SR, Fox EJ, *et al.* Oxidative stress is not a major contributor to somatic mitochondrial DNA mutations. PLoS Genet 2014; 10(2): e1003974.
[http://dx.doi.org/10.1371/journal.pgen.1003974] [PMID: 24516391]

[22] Hekimi S, Lapointe J, Wen Y. Taking a "good" look at free radicals in the aging process. Trends Cell Biol 2011; 21(10): 569-76.
[http://dx.doi.org/10.1016/j.tcb.2011.06.008] [PMID: 21824781]

[23] Madsen CS, Ghivizzani SC, Hauswirth WW. *In vivo* and *in vitro* evidence for slipped mispairing in mammalian mitochondria. Proc Natl Acad Sci USA 1993; 90(16): 7671-5.
[http://dx.doi.org/10.1073/pnas.90.16.7671] [PMID: 8356068]

[24] Mita S, Rizzuto R, Moraes CT, *et al.* Recombination *via* flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. Nucleic Acids Res 1990; 18(3): 561-7.
[http://dx.doi.org/10.1093/nar/18.3.561] [PMID: 2308845]

[25] Li M, Schröder R, Ni S, Madea B, Stoneking M. Extensive tissue-related and allele-related mtDNA

heteroplasmy suggests positive selection for somatic mutations. Proc Natl Acad Sci USA 2015; 112(8): 2491-6.
[http://dx.doi.org/10.1073/pnas.1419651112] [PMID: 25675502]

[26] Ramos A, Santos C, Alvarez L, Nogués R, Aluja MP. Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. Electrophoresis 2009; 30(9): 1587-93.
[http://dx.doi.org/10.1002/elps.200800601] [PMID: 19350543]

[27] Ramos A, Santos C, Barbena E, *et al.* Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the New Human Genome Reference Sequence (GRCh37). Electrophoresis 2011; 32(6-7): 782-3.
[http://dx.doi.org/10.1002/elps.201000583] [PMID: 21425173]

[28] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 1999; 23(2): 147.
[http://dx.doi.org/10.1038/13779] [PMID: 10508508]

[29] Santos C, Montiel R, Sierra B, *et al.* Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). Mol Biol Evol 2005; 22(6): 1490-505.
[http://dx.doi.org/10.1093/molbev/msi141] [PMID: 15814829]

[30] Spalding KL, Bhardwaj RD, Buchholz BA, Druid H, Frisén J. Retrospective birth dating of cells in humans. Cell 2005; 122(1): 133-43.
[http://dx.doi.org/10.1016/j.cell.2005.04.028] [PMID: 16009139]

[31] Li B, Krishnan VG, Mort ME, *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009; 25(21): 2744-50.
[http://dx.doi.org/10.1093/bioinformatics/btp528] [PMID: 19734154]

[32] IBM Corp [Computer Program]. Version 21.0. Armonk (NY): IBM SPSS Statistics for Windows 2012.

[33] Andrew T, Calloway CD, Stuart S, *et al.* A twin study of mitochondrial DNA polymorphisms shows that heteroplasmy at multiple sites is associated with mtDNA variant 16093 but not with zygosity. PLoS One 2011; 6(8): e22332.
[http://dx.doi.org/10.1371/journal.pone.0022332] [PMID: 21857921]

[34] Ross JM, Stewart JB, Hagström E, *et al.* Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. Nature 2013; 501(7467): 412-5.
[http://dx.doi.org/10.1038/nature12474] [PMID: 23965628]

[35] Kirkwood TB. Evolution of ageing. Nature 1977; 270(5635): 301-4.
[http://dx.doi.org/10.1038/270301a0] [PMID: 593350]

[36] Kirkwood TB, Austad SN. Why do we age? Nature 2000; 408(6809): 233-8.
[http://dx.doi.org/10.1038/35041682] [PMID: 11089980]

[37] Medawar PB, Ed. An unsolved problem of biology. London: Lewis HK and Co. 1952.

[38] Larsson NG. Somatic mitochondrial DNA mutations in mammalian aging. Annu Rev Biochem 2010;

79: 683-706.
[http://dx.doi.org/10.1146/annurev-biochem-060408-093701] [PMID: 20350166]

CHAPTER 4

Human Y Chromosome Mutation Rate: Problems and Perspectives

Paolo Francalacci[*], **Daria Sanna** and **Antonella Useli**

Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy

Abstract: The information contained into the human genome is the result of a historic process involving all the ancestors of a specific human being and it is a powerful tool for reconstructing human evolution. Key information about human evolutionary history can be robustly assembled from variation of the two haploid segments at uniparental transmission: mitochondrial DNA and Y chromosome. These two genetic systems are not subject to the rearrangement of recombination, and are inherited linearly through generations, having mutation as the only possible source of variation. In particular, the Y chromosome has been extensively studied for the study of human evolution. A major goal of the evolutionary research is not only to elucidate the pathways of the human peopling and the demographic changes that shaped the present populations, but also to date these events. For this aim, the recognition of a correct mutation rate is crucial. Genomic mutation rates can be estimated either by direct observation of mutations in present-day families (*de novo* mutation rate), by calibrating genetic variation against archaeological/historical records (evolutionary rate), or by using a sequence extracted from ancient human remains of known chronology. In order to test whether the same methodology could give consistent results when applied to different experimental contexts, we applied the evolutionary rate based upon archaeological evidence to two independent sets of data. Despite the striking difference in the absolute value of the substitution rate, the TMRCA of corresponding nodes in the phylogenetic trees obtained from the two databases are remarkably similar.

Keywords: 1000 genome project, Human evolution, Human population genetics, Molecular clock, Mutation rate, Next Generation Sequencing, Phylogenetic trees, Polymorphisms, Sardinian population, Single Nucleotide Y chromosome.

[*] **Corresponding author Paolo Francalacci:** Dipartimento di Scienze della Natura e del Territorio, Università di Sassari, 07100 Sassari, Italy; Tel/Fax: +39 079228631 / +39 079228665; E-mail: pfrancalacci@uniss.it

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

4.1. INTRODUCTION

The information contained into the human genome, encoding the project needed for the development, is the result of a historic process involving all the ancestors of a specific human being and it is a powerful tool for reconstructing the past event of human evolution. In fact, even the vast majority of the human genome is shared among all living individuals, the small fraction differentiating each of us is the result of the mutations that we have inherited until the last common ancestors.

These changes are genetic markers that individuate lineages and track back the pathway in time and space. For this reason, the study of the genetic variation is a fundamental tool for the knowledge of our evolutionary history.

The knowledge of the evolution of the human genome is strictly dependent on the availability of informative genetic markers and their relative coverage of genetic variation, sustaining the phylogenetic reconstruction. Following the pioneer studies on classical markers, such as blood groups and protein polymorphisms summarized in [1], modern advancement of molecular biology allows investigating directly the source of variation, the DNA. In recent years, cutting-edge genotyping technologies have enhanced the resolution of genome wide analyses by using hundreds of thousands (300K to 650K) of single nucleotide polymorphisms (SNPs) [2 - 4]. Presently, the development of more and more effective technologies of high throughput sequencing, coupled with a decrease of the analytical costs, allowed the detection of the complete variation of the human genome through resequencing, and numerous projects are now ongoing with this objective.

However, not all parts of the whole genome are suitable for population studies. In fact, even if the evolutionary story of one population is obviously the same, different portions of the genome tell us different stories. It is apparent that essential genes cannot change with a quick rate, since they can produce important alterations of the major metabolic functions, and so they cannot differentiate the various populations, being useful only for the reconstruction phylogenesis at higher taxonomic level. Genetic regions with lesser biological importance may vary at higher speed, allowing differentiating among species, populations or even individuals. The higher is the pace of variation the closer is the evolutionary event on which we may infer. Moreover, the genetic turnover among generations is also due to different mating strategies and can accelerate or slow down according to cultural and social structures. Consequently, any inference should be considered in the light of the specific genetic system studied and only the integration of data coming from different markers can give a reliable picture of human evolution. The independent inheritance of maternal and paternal chromosomes and the mechanism of intra-chromosomal recombination make of the individual genome as a puzzle of different contributions coming from past generations. However, the proportion of the contribution of a given ancestor halves in average at each generation, becoming rapidly negligible after a few centuries. As an example, the contribution of an ancestor lived only ten generations earlier is, approximately, one part over one thousand. Therefore, any modern individual is the resultant of a myriad of singular stories coming from various provenience, and reflects the complex demographic history of one or more geographic areas. The autosomal markers are particularly valuable for recognizing correspondence between genetic and geographic distances, and many methods have been developed over time to describe the autosomal variation at the geo-demographic level, such as the principal component analysis (PCA) [5] and similar principal component based algorithms. More recently Spatial Ancestry Analysis (SPA) and admixture based software, such as Admixture software [6], which are able, through a Bayesian approach, to dissect the overall variability in different strata reflecting past merging of different populations. Using the principle that recent admixture implies larger linkage blocks, eventually de-structured by recombination, Hellental *et al.* [7] proposed a method to date the admixture event from two or more different populations.

The stochastic processes and the large variances accompanying coalescence times for individual loci, even if constituted by large non-recombining DNA sequences such as the mitochondrial DNA (mtDNA) or the Male Specific region of the Y chromosome (MSY), limit their inferential value. Therefore, a multilocus approach would be preferable: using the data generated by such large scale sequencing projects, a full evaluation of genetic variation of the nuclear genome would be extremely informative. However, working with autosomes as an evolutionary marker, confounding factors like genetic recombination, gene conversion and natural selection complicate phylogenetic reconstruction. Instead, considering the lack of recombination and the low reversion/recurrence rates, key information about human evolutionary history can be robustly assembled from variation of the only two haploid segments of the human genome at uniparental transmission, as mtDNA and MSY. These two genetic systems are not subject to the rearrangement of recombination, and consequently inherited linearly identical through generations: mutation is the only possible source of variation and nucleotide changes accumulate over time in the molecule. If we consider the example mentioned above (*i.e.* a ten-generation genealogy), while the autosomal component of the genome comes from about a thousands of ancestors, the unilinear component is inherited invariantly (a part of the occurrence of newly produced mutations) from just a single individual of the genealogy, mtDNA from the maternal and MSY for the paternal side. For this reason, mtDNA and MSY may be inappropriate for describing the complex demography of a geographic area although they are excellent tools for tracking diachronic patterns of the human peopling [8, 9], particularly as markers of migration and ancient contact among populations [10, 11]. All the mtDNAs and MSYs of living humans descend by coalescence from a single female and male respectively. Their mutations in respect to the common ancestor accumulated along lineages, and can be used for tracking back their evolutionary history. However, the haploid status of these genetic systems makes mtDNA and MSY more sensitive to the effect of genetic drift, which may significantly skew the haplogroup frequencies. This is particularly true for the Y chromosome, considering the differences between male and female effective population sizes [12].

The first unilinear genome to be studied was in the late ’80s with the pivotal paper of Cann *et al.* [13], who provided a decisive support for the theory of the recent African origin of humanity, at that time still challenged by the multiregional theory [14]. In those years, the main molecular tool for evolutionary genetic analysis was the Polymerase Chain Reaction (PCR) allowing the amplification of small specific DNA segments. The very small mitochondrial genome (about 16,5 Kbp), present in many copies per cell and subject to high mutation rates due to the lack of proofreading activity of its polymerase, was an ideal model to study human evolution. Whereas the first studies based on Restriction Fragment Length Polymorphisms (RFLPs) [13] and on sequencing of the two hypervariable regions (HVS-I and II) [15], following studies combined the two approaches to get higher resolution [16]. Finally, with the improvement of cost effective sequencing techniques, the maximum extent of polymorphism coverage was reached for this genome, and the complete sequencing of mtDNA molecules is now routinely performed on a large-scale population level [17, 18].

4.2. THE Y CHROMOSOME

Because of its much larger molecular size, a similar approach has long been impractical for the Y chromosome (about 60 Mb for Y chromosome compared to 16.569 bp for mtDNA), but new sequencing technologies allow a comprehensive assessment of the entire genome, and several extensive efforts aim to catalogue genetic variation in different ethnicities, such as the 1,000 Genomes Project [http://www.1000genomes.org/]. Therefore, the whole variability of patrilinear genome represented by the MSY is presently accessible to the molecular analysis, giving to this system huge potentiality for reconstructing human evolution, at least from the paternal side.

However, before the full resequencing technology was available, the research on the Y chromosome knew historically different experimental methodologies. Most of the studies before the year 2000 were performed using Alu insertion [19] or STRs (Short Tandem Repeats) [20] with the known limitations due to recurrence and reversion of this kind of polymorphisms. Using Denaturing High Performance Liquid Chromatography (D-HPLC) technology, Underhill and co-workers [21] discovered 22 new SNP biallelic markers and few years later this method yielded 167 suitable for evolutionary analysis [22]. The first consensus phylogeny was published in 2002 by a consortium grouping many researchers from various laboratories (the Y Chromosome Consortium - YCC) [23], which adopted an alphanumeric nomenclature of the lineages modeled upon the one established for mtDNA. Updates of the tree were published some years later, growing the number of informative markers and the complexity of the topology, in 2003 [24] and 2008 [25]. Since then, several publications have been focused on specific clades of the tree, or devoted to population studies. A review for the European Y chromosomes and populations studies can be found in several papers [26, 27]. Recently, all published phylogenetic updates and Y-SNPs were combined in one phylogenetic tree [28], which reached the maximum complexity of the human Y chromosome phylogeny before the introduction of the new re-sequencing technologies. Starting from 2013, large-scale analyses of complete Y chromosome sequences (at least in the readable portion presently accessible) increased enormously the number of available markers with a consequent proportional impact on the Y phylogeny [29 - 31].

The complexity of the phylogeny is directly dependent from the number of available SNPs: from the early 242 of the YCC (2002), this number raised rapidly to about 600 [25], and up to more than 1,725 listed in the Y-DNA SNP Index 2012 of International Society of Genetic Genealogy (ISOGG) website (www.isogg.org). Finally, the new techniques based on full genome resequencing boasted the number of evolutionary informative SNPs of an order of magnitude. In 2010, the 1,000 Genomes Project Consortium announced the discovery of 2,780 SNPs in their pilot phase, based on low pass sequencing of the whole-genome of 77 males from four distinct populations. More recently, Wei *et al.* [31] analyzed at high coverage 8.97 Mb from the Y chromosome sequences of 35 samples released by Complete Genomics and one from the 1,000 Genome from worldwide populations and discovered 6,662 new variants. In 2013 a paper of Francalacci *et al.* [29] based on 1,200 Y chromosomes from the isolated Sardinian population raised the number of SNP markers to 11,763, and an extension of this work to the whole readable portion [32] of the chromosome reached the number of 22,240 informative SNPs. Combining the data from other projects [30, 33 - 35] the available SNP markers in the MSY are now in the order of more than 100,000.

The non-recombining MSY consists of about 56.4 Mbp, excluding the 3 millions of base pairs (Mbp) of the two telomeric pseudoautosomal regions (PAR) that recombine with the X chromosome (Fig. **1**). Only 23.1 Mbp have been mapped in the assembled human reference sequence (Hg 19, GRCh37), since the rest is made of repetitive constitutive heterochromatin in the centromere and in the long arm of the chromosome which is, for practical purposes, unreadable.

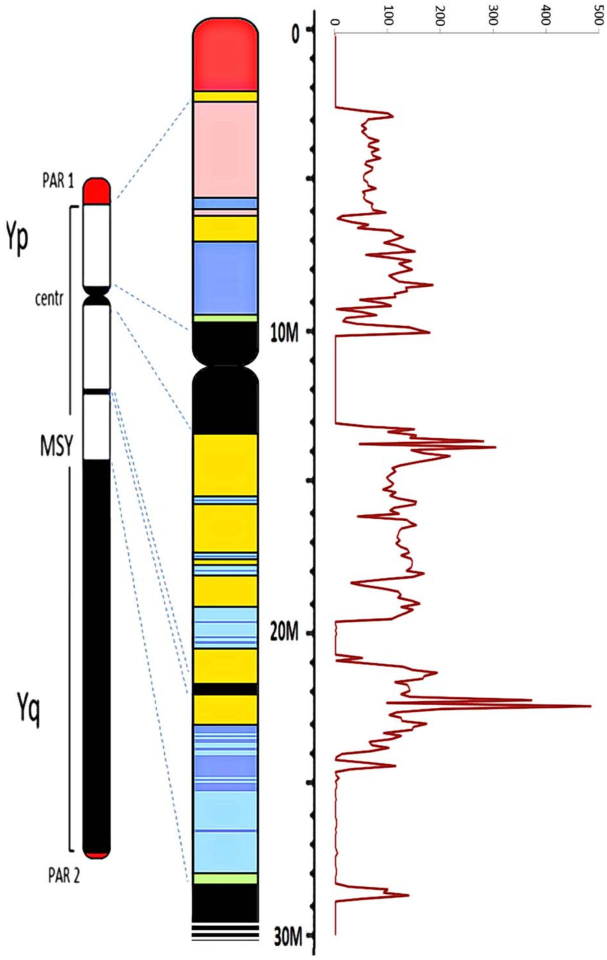


Fig. (1). Schematic representation of Y chromosome including the pseudoautosomal (red) and heterochromatic regions (black) and enlarged view of the euchromatic portion of the MSY with different classes of euchromatic sequences: X-degenerate (yellow), X-transposed (pink), Ampliconic (blue), palindromes within Ampliconic classes (light blue), and other (green). Redrawn from [32]. The red line indicates the distribution of the informative SNPs along the molecule.

The majority of the non-recombining euchromatic region falls into three classes [36]:

1. X-transposed sequences (3.4 Mbp), presenting a 99% homology to DNA

sequences in Xq21, as a result of a X-to-Y transposition, occurring after the divergence of the human and chimpanzee lineages;

2. Ampliconic sequences (9.7 Mbp), with a marked self-identity prone to gene conversion phenomena and exhibiting, in the long arm, eight palindromes and two inverted repeats with a 99.95% identity;
3. X-degenerate sequences (8.6 Mbp), with lower similarity with X chromosome and encompassing single-copy gene or pseudogene homologues of different X-linked genes.

In addition, about 0.4 Mbp of euchromatic sequences could not be classified in any of the three classes and were labelled as "Other".

Among the heterochromatic portions, about 1.0 Mbp of sequences, mainly located close to the centromere and in a small region interposed between two X-degenerate segments in the long arm, were sequenced in the last reference sequence release, raising the total amount of readable Y chromosome non-recombining sequences to about 23.1 Mbp.

Through resequencing, the MSY has now almost reached mtDNA in terms of knowledge of the total variability, at least for the readable portion, and presents several advantages for evolutionary analysis comparing to the mitochondrial genome. In fact, even if the mutation rate is lower, this is not a disadvantage since events like recurrent and reverse mutations, which confound the reconstruction of the mitochondrial molecular phylogeny, have a marginal impact on the Y chromosome. In addition, the much larger molecular size grants to the researcher a far larger number of phylogenetically informative markers. However, the two genetic systems, although sharing similar modality of inheritance, evolve differently, so any comparison between them in terms of reconstruction of the migratory patterns and their respective chronology should be considered with caution. In fact, differences in the sex ratio of migrant populations and in the effective population size (*i.e.* number of males *versus* females contributing to the next generation) may result both in a differential signal of gene flow of mtDNA *versus* MSY and in a unequal value of the Time of the Most Recent Common Ancestor (TMRCA) of males and females in a given population. An absolute difference between the TMRCAs of the MSY and mtDNA can in fact be expected because of randomness in the genealogical process. Under a model of constant population size the TMRCA varies widely among haploid loci. As each haploid locus only represents a realization of a stochastic process, the TMRCA of the MSY might significantly differ from that of the mtDNA.

4.3. DATING THE TREE

The major goal of the evolutionary research is not only to elucidate the pathways of the human peopling and the demographic changes that shaped the present populations but also to date these events. This is particularly interesting for the patrilineal MSY, whose present variation, because of the characteristics underlined above, coalesce through numerous mutational events to a single individual lived in the past carrying the most recent common ancestor of all the extant chromosomes. Therefore, the MSY variability, under the assumption of a constant substitution rate among lineages, reflects the origin or expansion of a specific Y lineage, allowing inferences on migration, gene follow, population split.

In earlier studies, the rather small number of available biallelic SNPs leaded to the classification of all human variation in a relatively low number lineages (or "haplotypes"), usually classified in families of lineages (or "haplogroups") characterized by one or more key mutations inherited from a common ancestor. In many cases, few haplotypes encompassed a very large number of individuals with little or no internal variability. Therefore, the poor variation of many haplogroups hampered the possibility of using biallelic SNPs as a marker for dissecting the chronology of the Y chromosome phylogeny. In order to overcome this problem, researchers used fast evolving markers, such as Short Tandem Repeats (STR) for giving an estimation of the chronology.

When applied to Y chromosome, the STR dating suggested that the common origin of all present human being is much more recent than previously believed, with a coalescent age of approximately 50,000 years [37, 38]. Moreover, both studies indicated the existence of a considerable population growth and an African origin for humankind. The newly proposed age is much younger than dates derived from mtDNA analysis, whose the coalescent age was ranged from about

150,000 [39] to 200,000 years ago [13].

However, although this approach is widely used, there is a current debate about the use STRs in MSY dating. In particular, the choice of the mutation rate, either calculated on pedigree, father-son pairs or calibrated on the base of historic event [40], is controversial [41]. A major problem of the use of Y-STR for the calculation of the mutation rate is that their variability appears to saturate rapidly. The use of networks analysis [42], especially if Y-SNPs were integrated into the analysis in order to separate the STR variation among haplogroups, could help to recognize the Y- STR mutational steps hindered by recurrent mutation. However, this strategy is not fully successful, and the use of observed mutation rates leads to time estimates that are several-fold too recent [43].

Similar results were also obtained when applied to the first discovered set of 167 SNPs, whose TMRCA was calculated at 59,000 years with a mutation rate of one new SNP every 6,900 years [22]. The lineages detected in this study were extremely non-homogeneous in terms of harbored variability, with a range of 2 to 13 SNPs from the common ancestor and, at a first sight, it might suggest the conclusion of a different mutation rate along time for each branch of the tree. Theoretically, such diversity could be related to factors like the genetic drift or purifying selection, favoring haplotypes showing certain specific mutations in respect of others, and thus altering a uniform accumulation of the variation over time in different lineages. The Y chromosome contains few but important genes on male fertility, and any mutational advantage in these genes, because of the lack of recombination would affect the entire chromosome and produce an increase in frequency of a lineage more rapidly than would be expected by drift [24]. This effect can be highlighted by association studies which measure the frequencies of Y haplotypes in matched groups of men with different phenotypes in terms of fertility. However, many studies showed no association: in other words, analyses of normal individuals showed that particular lineages have expanded more rapidly than others, having no relationship with any phenotype. The most famous case reported in literature [44] was a group of related haplotypes widespread in Central Asia in the approximate area of the Mongol Empire, but its distribution can be explained by social (possibly derived from the elite dominant family of Genghis Khan) rather than natural selection. Therefore, the evidence available so far does indicate that contemporary Y lineages can be assumed as neutral haploid markers [24]. Under the neutral theory, the rate of fixation of a neutral mutation is independent of the population size, hence not influenced by genetic drift, and it is equal to the mutation rate, which is the sole parameter influencing the rate of sequence evolution. This is the theoretical base for the application of the molecular clock principles, used for infer the timing of evolutionary events. Through the accumulation of neutral mutation, a new haplotype, as it expands, gives birth to a population of closely related haplotypes (*i.e.* sub-haplogroup): the amount of intra-allelic diversity among the haplotypes of this haplogroup is proportional to the age since they descend from a single common ancestor.

Since all extant chromosomes descend from a single common recent ancestor, they should harbor a similar number of neutral mutations for each haplotype, regardless of the frequency of the haplogroup to whom they belong in a specific geographic area, which is the result of drift, migration, expansion and other demographic events. Therefore, the high variance observed in the first set of 167 biallelic SNPs (116 haplotypes with 8.61+2.1 SNPs, range 2-13) [22], based only on a small number of markers, was likely biased by the randomness of the discovered variability. This random effect is expected to be reduced with the increase of the number of SNPs, which are now available in the order of thousands. In fact, a study based on a 2X sequence coverage of 1,200 Sardinian Y chromosomes detected 11,763 SNPs and the average length of each haplotype from the putative common ancestor was 1,002.6±21.2 [29]. This variance was even reduced increasing the sequence coverage (thus lowering the occurrence of lost markers) as in the work of Karmin *et al.* [35] carried at 30X in 454 individuals from 110 populations worldwide. The 37,475 SNPs were distributed among haplotypes with and average number of derivate alleles of 1,905.4±21.6 SNPs, pointing to a remarkable uniformity of the branch length, in agreement with the neutral nature of these polymorphisms on the Y chromosome. The branch length uniformity observed in phylogenies at high resolution is consistent with a constant phylogenetic rate of the SNPs in different lineages over time and it can be effectively used as a molecular clock for the dating of branch points. Specific algorithms have been proposed for inferring the date of the most recent common ancestor for a given bifurcation between two lineages, such as the rho [45] or the

BEAST [46] statistics, but, for this purpose, the recognition of a correct mutation rate is crucial.

4.4. MUTATION RATES

Genomic mutation rates can be estimated either by direct observation of mutations in present-day families (*de novo* mutation rate) or by calibrating genetic variation against archaeological/historical records (evolutionary rate). A third approach, somewhat in between the two methods, has recently been developed following the technological improvement of ancient DNA (aDNA: DNA extracted from bones or other tissues of extinct species or individuals) analysis, and consists in using a sequence extracted from an ancient human remain of known chronology. All the three methodologies require a constant mutation rate over time, an assumption that is not necessarily true in human evolution. Each one presents advantages compared to the others, but also drawbacks. All the mutation rate values given below are intended "per site per year" ($bp^{-1}y^{-1}$), and are summarized in Table **1**.

Table 1. Methodology used, number of Mb analysed, mutation rate in $bp^{-1}y^{-1}$ and TMRCA in thousands years of age, in some relevant studies on human Y chromosome.

Reference	Methodology	n° of Mb	Mutation Rate	TMRCA
Thomson *et al.* 2000 [37]	chimp/human	0.064	1.24×10^{-9}	50
Kuroki *et al.* 2006 [62]	chimp/human	13	1.50×10^{-9}	n.a
Xue *et al.* 2009 [49]	pedigree	10.15	1.00×10^{-9}	n.a
Mendez *et al.* 2013 [33]	pedigree	0.24	0.62×10^{-9}	338
Francalacci *et al.* 2013 [29]	archaeology	8.97	0.53×10^{-9}	180-200
Poznik *et al.* 2013 [30]	archaeology	10	0.82×10^{-9}	120-156
Helgason *et al.* 2015 [50]	pedigree	23.1	0.87×10^{-9}	n.a
Karmin *et al.* 2015 [35]	aDNA	10.8	0.74×10^{-9}	150
Trombetta *et al.* 2015 [51]	aDNA	1.5	0.72×10^{-9}	291

4.4.1. *de novo* Mutation Rate

The pedigree-based mutation rate estimation, widely used in the biomedical research, takes into account the number of new (*de novo*) mutations in two or more related individuals, divided for the number of meiosis which separate them.

This has been calculated by analyzing the whole genome of family trios or quartet [47, 48]. It requires a sequencing technology able to detect all the new mutations occurred in the two individuals analyzed, in order to avoid underestimation of the mutation rate, while the value should be an average statistically significant of many pairs of individuals analyzed, to avoid the bias due to highly stochastic mutational event.

The main advantage of this methodology is that the mutation rate estimation is obtained in a direct way on a known pedigree but, on the other hand, the application of the rate to a specific dataset is usually indirect since it derives from an external source (*i.e.* the rate is found in literature and not directly calculated on the analyzed samples). The disadvantage of this methodology instead is that not all the genomic regions evolve with the same rate. Therefore, a *de novo* mutation rate based on autosomes may not be suitable to be applied at the Y chromosome. Even if calculated on Y chromosome (such as in Thomson *et al.* [37] with four Y chromosome genes in a 13-step pedigree and in Helgason *et al.* [50] resequencing 23.1 Mbp of Y chromosome in deep rooted Icelandic genealogies) the value could be misleading if applied to dataset in which the sequences encompass MSY regions different from the four genes of the original study. In fact, a recent study highlighted that different regions within the Y chromosome have different mutation rates [51], so even a Y chromosome specific rate is not necessarily suitable to be applied “as it is” to studies investigating different regions of MSY. That is to say, that any substitution rate can only be applied to that specific DNA region on which it has been calculated [51].

When the *de novo* mutation rate is used for evolutionary studies in order to obtain a chronological estimation, the mutation per generation should be converted into mutation per year. That conversion imposes another problematic issue: the correct estimation of the generation time, which consists in the average of the ages of a male between the first and the last son of his reproductive life. This is an extremely difficult parameter to estimate in archaeological context since it can be highly variable and strictly dependent on population dynamics. Nowadays, there is a general agreement for the age of 30 [49], although ethnological studies showed a wide variance depending on the social, economic and ethnic contexts. Even though the use of historic sources can extend the direct knowledge of the generation time from several centuries to almost one millennium, this parameter is virtually impossible to estimate in prehistoric time. The ancient generation time was very likely shorter than the current one, coherently with a presumably shorter duration of life. In addition, many studies have noted that molecular rates observed on genealogical timescales are greater than those measured in long-term evolutionary times, raising the problem about the suitability of applying a rate calculated on extant (or only few generations old) individuals to an evolutionary framework [52]. Moreover, the number of mutations observed in Y chromosome is strongly correlated with the age of the father [53], adding further importance to a precise, although difficult, estimation of the generation time.

This pedigree-based rate has been widely used in Y chromosome demographic and lineage dating, often using adjusted mutation rates either derived from autosomes or directly calculated in Y chromosome. The former approach was used by Mendez *et al.* [33], who considered the paternal autosomal mutation rates reported on an Icelandic dataset of 78 parent-offspring trios. The authors applied this rate (6.17×10^{-10} with 20-40 years of assumed generation time) to a very divergent Y chromosome lineage (named A00) of two African individuals from Cameroon, obtaining a TMRCA of 380,000, thus pre-dating the emergence of *H. sapiens* in Africa. However, Elhaik *et al.* [54] have challenged this extraordinary early date, criticizing, among others, the choice of an unreasonable generation time. The same rate was used also by Scozzari *et al.* [34] to date the bifurcation immediately downstream (named A1b, corresponding to A0 elsewhere) with a result of 196,000. This same research group had previously preferred the latter approach [55] to date the A0 split: the use of the rate calculated on Y chromosome [49] yielded a TMRCA of 142,000. This date is older than that previously calculated using STRs, but younger of about 50,000 years in respect to the one obtained applying the Mendez *et al.* [33] method.

4.4.2. Evolutionary Rate

In a given phylogeny, the evolutionary rate (also called phylogenetic rate) requires a split between two lineages which could be related to a chronologically known event. Consequently, such event will be used as a calibration point of the molecular clock, dividing the average number of mutations downstream with the number of years passed after the separation. Finally, the obtained rate will be used for dating all the other nodes of the tree. Therefore, the estimation in indirect, requiring evidence based on other fields such as Paleontology (*e.g.* human-chimpanzee separation [37]), Archaeology (*e.g.* peopling of the Americas [30]; peopling of Sardinia [29]), History (*e.g.* Maori invasion of Cook Islands [40]), *etc*. This approach is thus strongly dependent both on the precise dating of the calibration point, which is sometimes rather uncertain in archaeological context. In addition to that, it is also influenced by the occurrence of a strict association between the historic event and the molecular divergence since there may be a substantial time lag between a mutation and the demography at the base of the spread of the sampled chromosome that carries it [56]. This limitation becomes almost irrelevant when a large number of mutations with a densely distributed occurrence over time are present since, in this case, the time lag would be minimal. Furthermore, the sequences taken into account for setting the calibration point are encompassed in the dataset, thus sharing all the features of the other sequences in term of coverage, region studied, *etc*. Therefore, the evolutionary rate is not biased by the choice of the chromosomal region to be studied and can be applied even in case of incomplete knowledge of the variability, if the missing data is homogeneously distributed through individual samples. Another significant advantage is that the conversion in mutation per years can be done without establishing a predetermined generation time, thus overcoming this important source of error. It should also be considered that, if the evolutionary rate is based on an incomplete variability extraction, it does not represent an absolute value and it is expected to vary in case that more variability is further discovered. Thus, being a relative rate, its application is limited to the phylogenetic tree on which it has been calculated. However, even if two distinct phylogenies with different resolution in terms of variant discovery show divergent rates, those two rates, if correctly applied, should give the same chronology to analogous bifurcations of the trees. Moreover, the congruence of the reconstructed chronologies constitute a support for the correctness of the calibration point.

The archaeologically calibrated molecular clock methodology was followed by Poznik *et al.* [30], who used the putative period of colonization of the Americas (estimated as 15,000 years ago) to date the split between the two main branches of the Amerindian haplogroup Q, and by Francalacci *et al.* [29], who considered the expansion of the Mediterranean Sardinian population observed in early Neolithic period around 7,700 years ago. The mutational rates calculated in the two studies were rather different, being 0.72x10^{-9} for the former and 0.53x10^{-9} for the latter, but as explained above, they cannot be considered as absolute values since they are dependent on the discovered variability. A higher coverage (especially for Francalacci *et al.* [29], who used a low-pass analysis at 2X coverage) would have yielded additional SNPs and consequently changing the evolutionary rate.

4.4.3. Ancient DNA Based Rate

This methodology involves the sequencing of Y chromosomes from ancient samples for which reliable radiocarbon dates are available, used as a calibration point of the phylogenetic tree. Analogously to what has been done for mtDNA [57], it has been applied to Y chromosome in Karmin *et al.* [35] and Trombetta *et al.* [51], even if the analysis of an ancient specimen was already carried out on Francalacci *et al.* [29] with a validation purpose. Recently, the method has been used for dating the divergence between Neandertal and Modern Human Y chromosomes at 588,000 years ago [58]. This approach combines the advantages of the above methods: a direct estimation of the mutation rate (*i.e.* not relying on archaeologically based demographic inference, which is not always certain) as in the *de novo* method and a possibility of processing the ancient sequence together with the entire studied database (*i.e.* the value is not dependent on the chromosome regions analysed) as in the evolutionary method. Furthermore, it does not require the estimation of the generation time. However, it still has to overcome many difficulties, such as possible contamination or, chiefly, problems of low coverage since the ancient tissues are expected to yield a poor DNA quality with the concrete risk of an underestimation of their variability. However, with the fast emerging and growing of modern aDNA analytical techniques, entirely sequenced Y chromosomes in ancient individuals are now available at a reasonable coverage, with a concrete possibility of future improvement in the extraction techniques. Thus, this already promising methodology will probably become the elective one for the correct assessment of Y chromosome mutation rate [52]. Notwithstanding, if appropriately adapted in terms of congruence with the modern database, the mutation rates obtained using ancient specimens are already quite consistent. In fact, the rate obtained by Karmin *et al.* [35], using two ancient Amerindian skeletons (Anzick, dated 12,600 years ago, and Saqqaq, dated 40,00 years ago) for calibration was 0,74x10^{-9}, matching the one calculated by Trombetta *et al.* [51], who considered the ancient remains of Ust'-Ishim (45,000 years ago) and Loschbour (60,000 years ago) and resulted 0,72x10^{-9}.

4.5. APPLICATION TO TWO PUBLIC DATABASES

Presently, the choice of which kind of mutation rate to use in Y chromosome dating is still controversial since different rates can result in temporal estimates that deviate several-fold (for a review see Wang *et al.* [52]). Nevertheless, in this field, research improvements should eventually lead to a convergence of the different analyses to a consensus mutation rate, and ultimately, to a coherent estimation of the chronology of the main demographic events of human evolutionary history. In order to test if the same methodology could give consistent results when applied to different experimental contexts, we used the evolutionary rate based upon archaeological evidence to two sets of data: the Sardinian database (Sdb) and the 1,000 Genomes database (1KGdb). The variable SNPs of the Sdb are available in Francalacci *et al.* [32] as supplemental material, while Y chromosome sequences of the phase 3 of the 1KGdb can be downloaded at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The 1KGdb is publicly available but, in agreement with the Fort Lauderdale principles, the data can be only used for comparative purposes until the data producers publish their first major paper. Therefore, we will only present here the aggregate general data to be compared with the Sdb and no details will be provided. The data were used for reconstructing the phylogeny of all the Y sequences of the two databases. The maximum parsimony tree, calculated using PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) [59] and drawn using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/), is shown in Fig. (**2**) for the Sdb, and in Fig. (**3**) for the 1KGdb.

The Sdb is constituted by 1,197 Sardinians and 7 non-Sardinian individuals, analyzed at a coverage of 2X for the whole readable portion of the Y chromosome, encompassing about 23.1 Mbp [32]. The samples yielded 22,240 SNPs including singletons and the average length of each haplotype from the putative common ancestor was 1,576.1±68.4. The analytical approach, based on low-pass sequencing at 2X, yielded an amount of variation that is an underestimate of the total one, and for this reason, an external *de novo* mutation rate would not be suitable. On the other hand, the peopling of an island, subject to bottlenecks and founder effects, offers a great opportunity to establish the amount of variability produced after the settlement, whereas a considerable number of lineages results private to the island, providing an excellent calibration point suitable for an evolutionary rate calculation. The archaeological evidence indicates a putative age of demographic expansion about 7,700 years ago, coincident with the initial Neolithic period on the island. Among the private Sardinian clades, the ones that present the higher average branch length encompass 60 to 70 SNPs. The most representative of them, belonging to I2a1 sub-haplogroup downstream a non-Sardinian sample, includes a very large number of samples (430 individuals), with a large number of clusters joined by very short branches, suggestive of an expansion event [56]. Taking into account that the average variability of this private Sardinian clade is 66.2±13.3 SNPs, a calibration point of 7,700 years ago results in a provisional measure of the evolutionary rate of one new mutation every 116.3±20 years. Multiple founders with prior variability may result in a coalescent age that predate or postdate the colonization event, so that the last upstream shared SNP and the first downstream private SNP represent the upper and lower bound of a given demographic change. However, the accuracy of the estimate depends on the fixation rate over the time, and a fast evolving marker, such as the MSY studied at a resequencing level, could circumscribe the chronological gap: since in this phylogeny a new mutation is produced, in average, almost every century, a possible bias appears to be rather limited.

Calculating approximately 23.1 Mbp sequenced for each Y chromosome, our phylogenetic rate is equivalent to about $0.37x10^{-9}$, which is remarkably lower than either the archaeologically [29, 30] or pedigree [31, 50] based estimates previously reported. It is worth noting that it does not represent the absolute mutation rate but it is strictly dependent on the fraction of the actual informative variation extracted with a low pass sequencing approach, and, as said above, can be used only for dating the phylogeny on which it has been calculated. Since the density of phylogenetically informative SNPs outside the X-degenerate class is poorer (possibly due to technical difficulties in the variant calling procedure), the phylogenetic rate of the entire sequenced MSY is about half of the one previously calculated on mainly X-degenerate sequences using the same database [29]. Applying this phylogenetic rate, the TMRCA of all samples is estimated around 183,000 years ago (with a confidence interval of 191,000-175,000 years). Despite of the striking difference in the absolute value of the substitution rate, it is worth stressing that the TMRCA calculated on the complete 23.1 Mbp portion is identical to the one reported in Francalacci *et al.* [29].

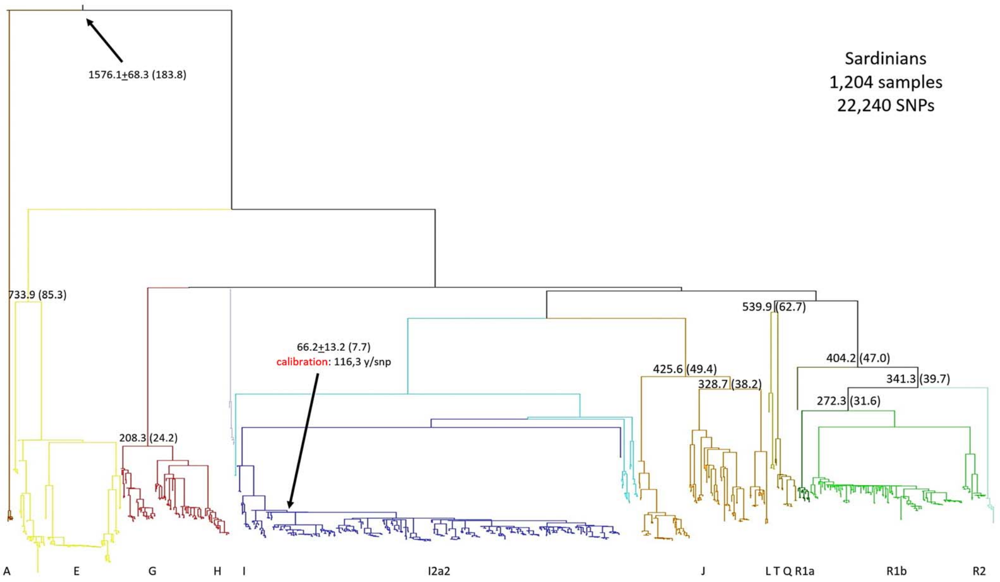


Fig. (2). Phylogenetic tree of the 1194 Sardinians Y-chromosome sequences. Colored branches represent different Y-chromosome haplotypes. The arrows indicate the calibration point and the most recent common ancestor of the tree. At each relevant node are reported the average number of downstream SNPs and, in brackets, the putative time of origin in thousands of years.

The 1KGdb is constituted by 1,233 individuals from populations from all continents, analysed at low coverage for the X-degenerate region of the MSY, encompassing about 9.1 Mb. The samples yielded 59,274 SNPs including singletons and the average length of each haplotype from the putative common ancestor was 1,393.9 ±30.5 SNPs. To calibrate the molecular clock, we considered the archaeological calibration reported in Poznik *et al.* [30], who relied on well-dated archaeological sites, which indicate that humans first colonized the Americas about 15,000 years ago, and assumed that the two main divergent private Amerindian sub-haplogroups Q-M3 and Q-L54*(xM3) diverged at about the same time with the initial peopling of Americas. Applying this calibration point to the 1KGdb, and taking into account that the average variability of the two Native American clades is 102.0±20.5 SNPs, a calibration point of 15,000 years ago results in a provisional measure of the evolutionary rate of one new mutation every 147.0±20 years. Considering that these variants have been found on the X-degenerate region of the MSY, equivalent to about 9.1 Mb, the evolutionary rate is 0.75×10^{-9}. Applying this evolutionary mutation rate, the TMRCA of the whole sample, which includes the A0 haplogroup, goes back to 205,000 years ago, and 190,000 years ago (with a confidence interval of 195,000-185,000 years) if we consider the A1b node, equivalent to the one present in the Sdb.

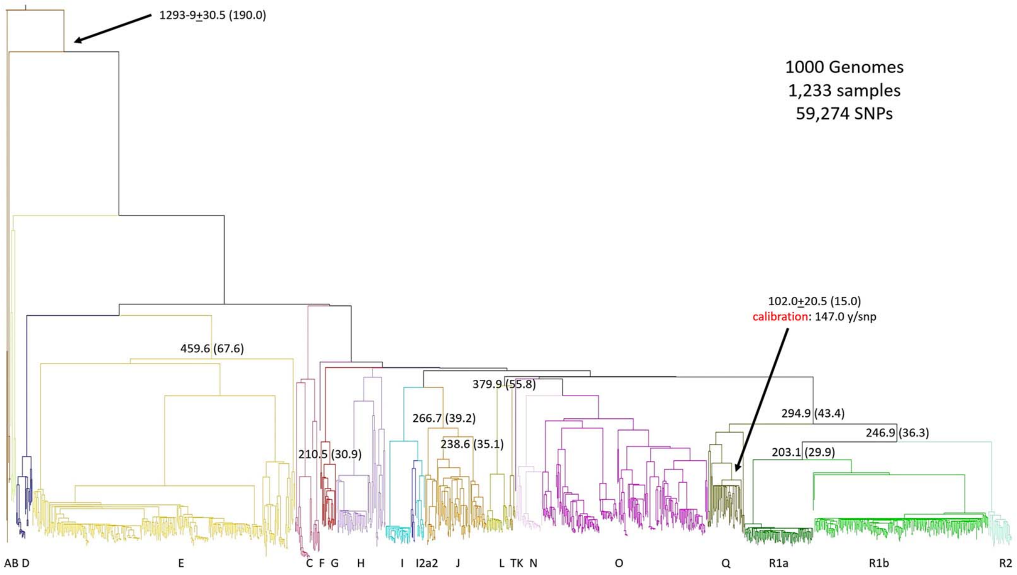


Fig. (3). Phylogenetic tree of the 1,223 1,000Genomes Y-chromosome sequences. Colored branches represent different Y-chromosome haplotypes. The arrows indicate the calibration point and the most recent common ancestor of the tree. At each relevant node are reported the average number of downstream SNPs and, in brackets, the putative time of origin in thousands of years.

4.5.1. Comparing the Two Databases

The two databases, encompassing a similar number of independent samples analyzed at low coverage, show apparent differences both in terms of total number of recovered SNPs and relative evolutionary rate.

The total number of SNPs is doubled in the 1KGdb in respect to the Sdb, even if the MSY portion analyzed is much smaller, as limited to the X-degenerate for the former and extended to the whole readable portion for the latter. This discrepancy is easily explainable considering the sample composition, which is encompassed by individuals coming from worldwide populations in the 1KGdb and restricted to a single Euro-Mediterranean population for the Sdb, thus lacking most of the haplogroups which are non-typically European. The average branch length is rather similar in the two databases, being slightly shorter in the 1KGdb, which is however based on a much smaller portion of the MSY.

The evolutionary rate calculated in the Sdb is approximately half of the one determined in the 1KGdb and, in general, very low, while the rate estimated in the 1KGdb is similar to other rates recently proposed, based on the X degenerate portion and obtained including ancient remains [35, 51].

Despite of the striking difference in the absolute value of the substitution rate, applying two independent calibration points (the Sardinian Neolithic expansion of Sdb and the peopling of the Americas for the 1KGdb) the TMRCAs of corresponding nodes are remarkably similar in the phylogenetic trees obtained from the two databases. The general TMRCA is comprised in the range of 200-180,000 years ago and in particular, the separation of the A1 clade from the rest (a node in common between the two databases) of the tree is almost coincident. This TMRCA is consistently older than the one generally accepted using STR variation, although is in full agreement with the revised molecular clock for humans [60], proposed in accordance with the fossil evidence of modern human in Africa about 200,000 years ago [61]. In addition, it is consistent with the TMRCA estimated from analyses of maternally inherited mtDNA [13, 39], for which complete sequencing provides a full catalogue of variation (although, as pointed out above, this coincidence might not be necessarily obligatory for the two independent haploid genomes). Also, the origin of the main European haplogroups whose correspondences can be found in the two phylogenetic trees (such as the origin of haplogroup E, G, J, P, R or sub-haplogroups J2 or R1a) are consistently similar in the two independent datasets.

CONCLUSIVE REMARKS

The evolutionary rate, although relative to the dataset on which it has been calculated, can produce reliable results among different samples. Moreover, the convergence of paleoanthropologic record and mtDNA analysis with our Y chromosome estimate suggests the possibility of using this TMRCA as a reference point for further studies whereas no other internal calibration is available. In this way, the problems linked to the *de novo* rate (such as the necessity of the establishment of a generation time) and those based on ancient remains (such as the contamination and coverage problems) could be overcome.

We are confident that technical improvements will soon provide convergence of the three different methodologies towards a consensus in the mutation rate and, consequently, in the timing of the Y chromosome evolutionary tree.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

We are grateful to all Sardinian donors for providing blood samples. We also thank the 1000 Genomes Project consortium for making available their sequencing data, in compliance with the Fort Lauderdale principles.

REFERENCES

[1] Cavalli-Sforza LL, Menozzi P, Piazza A. The History and Geography of Human Genes. Princeton, USA: Princeton University Press 1994.

[2] Tian C, Plenge RM, Ransom M, *et al.* Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 2008; 4(1): e4.

[3] Novembre J, Johnson T, Bryc K, *et al.* Genes mirror geography within Europe. Nature 2008; 456(7218): 98-101.
[http://dx.doi.org/10.1038/nature07331] [PMID: 18758442]

[4] Li JZ, Absher DM, Tang H, *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008; 319(5866): 1100-4.

[http://dx.doi.org/10.1126/science.1153717] [PMID: 18292342]

[5] Elhaik E, Tatarinova T, Chebotarev D, *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat Commun 2014; 5: 3513.
[http://dx.doi.org/10.1038/ncomms4513] [PMID: 24781250]

[6] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009; 19(9): 1655-64.
[http://dx.doi.org/10.1101/gr.094052.109] [PMID: 19648217]

[7] Hellenthal G, Busby GB, Band G, *et al.* A genetic atlas of human admixture history. Science 2014; 343(6172): 747-51.
[http://dx.doi.org/10.1126/science.1243518] [PMID: 24531965]

[8] Raghavan M, Skoglund P, Graf KE, *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 2014; 505(7481): 87-91.
[http://dx.doi.org/10.1038/nature12736] [PMID: 24256729]

[9] Achilli A, Perego UA, Lancioni H, *et al.* Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc Natl Acad Sci USA 2013; 110(35): 14308-13.
[http://dx.doi.org/10.1073/pnas.1306290110] [PMID: 23940335]

[10] Sikora MJ, Colonna V, Xue Y, Tyler-Smith C. Modeling the contrasting Neolithic male lineage expansions in Europe and Africa. Investig Genet 2013; 4(1): 25.
[http://dx.doi.org/10.1186/2041-2223-4-25] [PMID: 24262073]

[11] Hammer MF, Karafet TM, Park H, *et al.* Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet 2006; 51(1): 47-58.
[http://dx.doi.org/10.1007/s10038-005-0322-0] [PMID: 16328082]

[12] Jorde LB, Watkins WS, Bamshad MJ, *et al.* The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 2000; 66(3): 979-88.
[http://dx.doi.org/10.1086/302825] [PMID: 10712212]

[13] Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature 1987; 325(6099): 31-6.
[http://dx.doi.org/10.1038/325031a0] [PMID: 3025745]

[14] Wolpoff MH, Wu XZ, Thorne AG. Modern homo sapiens origins: A general theory of hominid evolution involving the fossil evidence from East Asia The origins of modern humans. New York, USA: Liss 1984; pp. 411-83.

[15] Torroni A, Schurr TG, Cabell MF, *et al.* Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 1993; 53(3): 563-90.
[PMID: 7688932]

[16] Torroni A, Huoponen K, Francalacci P, *et al.* Classification of European mitochondrial DNA from an analysis of 3 European populations. Genetics 1996; 144: 1835-50.
[PMID: 8978068]

[17] Achilli A, Rengo C, Magri C, *et al.* The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 2004; 75(5): 910-8.

[http://dx.doi.org/10.1086/425590] [PMID: 15382008]

[18] Pala M, Achilli A, Olivieri A, *et al.* Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 2009; 84(6): 814-21.
[http://dx.doi.org/10.1016/j.ajhg.2009.05.004] [PMID: 19500771]

[19] Hammer MF. A recent common ancestry for human Y chromosomes. Nature 1995; 378(6555): 376-8.
[http://dx.doi.org/10.1038/378376a0] [PMID: 7477371]

[20] de Knijff P, Kayser M, Caglià A, *et al.* Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 1997; 110(3): 134-49.
[http://dx.doi.org/10.1007/s004140050052] [PMID: 9228564]

[21] Underhill PA, Jin L, Lin AA, *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 1997; 7(10): 996-1005.
[PMID: 9331370]

[22] Underhill PA, Shen P, Lin AA, *et al.* Y chromosome sequence variation and the history of human populations. Nat Genet 2000; 26(3): 358-61.
[http://dx.doi.org/10.1038/81685] [PMID: 11062480]

[23] A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 2002; 12(2): 339-48.
[http://dx.doi.org/10.1101/gr.217602] [PMID: 11827954]

[24] Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 2003; 4(8): 598-612.
[http://dx.doi.org/10.1038/nrg1124] [PMID: 12897772]

[25] Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 2008; 18(5): 830-8.
[http://dx.doi.org/10.1101/gr.7172008] [PMID: 18385274]

[26] Francalacci P, Sanna D. History and geography of human Y-chromosome in Europe: a SNP perspective. J Anthropol Sci 2008; 86: 59-89.
[PMID: 19934469]

[27] Francalacci P, Morelli L, Useli A, Sanna D. The history and geography of the Y chromosome SNPs in Europe: an update. J Anthropol Sci 2010; 88: 207-14.
[PMID: 20834059]

[28] Van Geystelen A, Decorte R, Larmuseau MHD. Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs. For Sci Int 2013; 7: 573-80.

[29] Francalacci P, Morelli L, Angius A, *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 2013; 341(6145): 565-9.
[http://dx.doi.org/10.1126/science.1237947] [PMID: 23908240]

[30] Poznik GD, Henn BM, Yee MC, *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males *versus* females. Science 2013; 341(6145): 562-5.
[http://dx.doi.org/10.1126/science.1237619] [PMID: 23908239]

[31] Wei W, Ayub Q, Chen Y, *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 2013; 23(2): 388-95.
[http://dx.doi.org/10.1101/gr.143198.112] [PMID: 23038768]

[32] Francalacci P, Sanna D, Useli A, *et al.* Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human Y chromosome from a Sardinian sample. BMC Res Notes 2015; 8: 174.
[http://dx.doi.org/10.1186/s13104-015-1130-z] [PMID: 25926048]

[33] Mendez FL, Krahn T, Schrack B, *et al.* An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 2013; 92(3): 454-9.
[http://dx.doi.org/10.1016/j.ajhg.2013.02.002] [PMID: 23453668]

[34] Scozzari R, Massaia A, Trombetta B, *et al.* An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 2014; 24(3): 535-44.
[http://dx.doi.org/10.1101/gr.160788.113] [PMID: 24395829]

[35] Karmin M, Saag L, Vicente M, *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res 2015; 25(4): 459-66.
[http://dx.doi.org/10.1101/gr.186684.114] [PMID: 25770088]

[36] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 2003; 423(6942): 825-37.
[http://dx.doi.org/10.1038/nature01722] [PMID: 12815422]

[37] Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci USA 2000; 97(13): 7360-5.
[http://dx.doi.org/10.1073/pnas.97.13.7360] [PMID: 10861004]

[38] Shen P, Wang F, Underhill PA, *et al.* Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 2000; 97(13): 7354-9.
[http://dx.doi.org/10.1073/pnas.97.13.7354] [PMID: 10861003]

[39] Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 1995; 92(2): 532-6.
[http://dx.doi.org/10.1073/pnas.92.2.532] [PMID: 7530363]

[40] Zhivotovsky LA, Underhill PA, Cinnioğlu C, *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 2004; 74(1): 50-61.
[http://dx.doi.org/10.1086/380911] [PMID: 14691732]

[41] Wang CC, Li H. Discovery of phylogenetic relevant Y-chromosome variants in 1000 genomes project data. Preprint arXiv 2013. 1310.6590.

[42] Bandelt H-J, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999; 16(1): 37-48.
[http://dx.doi.org/10.1093/oxfordjournals.molbev.a026036] [PMID: 10331250]

[43] Wei W, Ayub Q, Xue Y, Tyler-Smith C. A comparison of Y-chromosomal lineage dating using either

resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci Int Genet 2013; 7(6): 568-72.
[http://dx.doi.org/10.1016/j.fsigen.2013.03.014] [PMID: 23768990]

[44] Zerjal T, Xue Y, Bertorelle G, *et al.* The genetic legacy of the Mongols. Am J Hum Genet 2003; 72(3): 717-21.
[http://dx.doi.org/10.1086/367774] [PMID: 12592608]

[45] Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 1996; 59(4): 935-45.
[PMID: 8808611]

[46] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007; 7: 214.
[http://dx.doi.org/10.1186/1471-2148-7-214] [PMID: 17996036]

[47] Awadalla P, Gauthier J, Myers RA, *et al.* Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. Am J Hum Genet 2010; 87(3): 316-24.
[http://dx.doi.org/10.1016/j.ajhg.2010.07.019] [PMID: 20797689]

[48] Roach JC, Glusman G, Smit AF, *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 2010; 328(5978): 636-9.
[http://dx.doi.org/10.1126/science.1186802] [PMID: 20220176]

[49] Xue Y, Wang Q, Long Q, *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 2009; 19(17): 1453-7.
[http://dx.doi.org/10.1016/j.cub.2009.07.032] [PMID: 19716302]

[50] Helgason A, Einarsson AW, Guðmundsdóttir VB, *et al.* The Y-chromosome point mutation rate in humans. Nat Genet 2015; 47(5): 453-7.
[http://dx.doi.org/10.1038/ng.3171] [PMID: 25807285]

[51] Trombetta B, D'Atanasio E, Massaia A, *et al.* Regional differences in the accumulation of SNPs on the male-specific portion of the human Y chromosome replicate autosomal patterns: Implications for genetic dating. PLoS One 2015; 10(7): e0134646.
[http://dx.doi.org/10.1371/journal.pone.0134646] [PMID: 26226630]

[52] Wang CC, Gilbert MT, Jin L, Li H. Evaluating the Y chromosomal timescale in human demographic and lineage dating. Investig Genet 2014; 5: 12.
[http://dx.doi.org/10.1186/2041-2223-5-12] [PMID: 25215184]

[53] Kong A, Frigge ML, Masson G, *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. Nature 2012; 488(7412): 471-5.
[http://dx.doi.org/10.1038/nature11396] [PMID: 22914163]

[54] Elhaik E, Tatarinova TV, Klyosov AA, Graur D. The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. Eur J Hum Genet 2014; 22(9): 1111-6.
[http://dx.doi.org/10.1038/ejhg.2013.303] [PMID: 24448544]

[55] Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. Am J Hum Genet 2011; 88(6): 814-8.

[http://dx.doi.org/10.1016/j.ajhg.2011.05.002] [PMID: 21601174]

[56] Jobling M, Hurles ME, Tyler-Smith C. Human evolutionary genetics. Garland Science, New York, USA: Origins, Peoples and Disease 2004.

[57] Fu Q, Mittnik A, Johnson PL, *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 2013; 23(7): 553-9. [http://dx.doi.org/10.1016/j.cub.2013.02.044] [PMID: 23523248]

[58] Mendez FL, Poznik GD, Castellano S, Bustamante CD. The divergence of neandertal and modern human Y chromosomes. Am J Hum Genet 2016; 98(4): 728-34. [http://dx.doi.org/10.1016/j.ajhg.2016.02.023] [PMID: 27058445]

[59] Felsenstein J. PHYLIP - phylogeny inference package (Version 3.2). Cladistics 1989; 5: 164-6.

[60] Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 2012; 13(10): 745-53. [http://dx.doi.org/10.1038/nrg3295] [PMID: 22965354]

[61] Gibbons A. Human evolution. Turning back the clock: slowing the pace of prehistory. Science 2012; 338(6104): 189-91. [http://dx.doi.org/10.1126/science.338.6104.189] [PMID: 23066056]

[62] Kuroki Y, Toyoda A, Noguchi H, *et al.* Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nat Genet 2006; 38(2): 158-67. [http://dx.doi.org/10.1038/ng1729] [PMID: 16388311]

CHAPTER 5

Genomics of Isolated Populations: Inferences for Gene-Finding Studies

Manuela Lima[1,2,3,*]

[1] *Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal*

[2] *Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal*

[3] *Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal*

Abstract: Genetic isolates correspond to subpopulations that have derived from a reduced number of individuals originally belonging to a main parental population, which have been submitted to isolation. Such subpopulations are widely acknowledged as important resources for the elucidation of the genetic basis of diseases. Mendelian diseases in particular have profited from gene finding strategies that use genetically isolated populations. In view of the success obtained for monogenic disorders, the scientific community was encouraged to use isolated populations with the purpose of analyzing complex diseases. In this chapter, the characteristics of human genetic isolates are addressed and the way by which they provide an advantage to gene finding studies is discussed. The implications of genomics for the efforts of gene identification using genetic isolates are also analyzed, and selected examples are provided. The availability of genomic data for isolated populations is currently providing in-depth insights into their structure, potentiating the use of research designs which are particularly suited for each isolate, thus increasing the chances of success of gene identification studies.

Keywords: Association studies, Complex diseases, Consanguinity, Endogamy, Genetic homogeneity, Human genetic isolates, Linkage disequilibrium Monogenic diseases, Non-extreme genetic isolates.

[*] **Corresponding author Manuela Lima:** Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal; Tel/Fax: 351 296650118; E-mail: maria.mm.lima@uac.pt

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

5.1. GENETIC VARIATION AND DISEASE IN HUMAN POPULATIONS

One main goal of human genetics research is the elucidation of the genetic basis of pathologies, a goal which implies the understanding of how genetic variation contributes to the biological pathways underlying the disease status [1]. Besides providing the possibility of molecular testing, with important implications for patients and families, gene identification allows the understanding of the pathogenic processes behind genetic disease, offering opportunities for therapeutic targets identification [2]. For monogenic disorders this has been a rewarding task, with nearly 4900 diseases with phenotype description and known molecular basis [3]. The scenario is clearly distinct for complex diseases, otherwise called "diseases of complex genetics", which arise from interactions between variants at several loci and the environment. Depending on a less clear genetic influence, this type of disorder has shown slow progresses on what concerns the elucidation of the underlying genetic basis: phenocopies, genetic heterogeneity, variable clinical expression, incomplete penetrance and environmental influences are some of the key factors which complicate the understanding of the genetic component of diseases [4]. In complex disorders, in addition to environmental factors, the potential presence of a large number of genes on the basis of the phenotype, each with a reduced contribution to the relative risk, further complicates the understanding of the underlying genetic basis [5].

In attempting to access the genetic component of a monogenic condition, linkage analysis has been considered the most powerful methodology. Linkage is dependent on the physical distance between loci, reflected in the recombination fraction, which is measured from the co-segregation observed in pedigrees of the trait of interest with genetic markers. The standard linkage methodology requires genotyping a large number of markers in several members (affected and unaffected) of a set of families. Limitations derived from the small number of available markers initially constituted crucial obstacles to gene identification studies, given the impossibility of adequately following the transmission of the phenotype. By the eighties of last century the growing availability of DNA markers prompted several attempts to localize genes responsible for Mendelian diseases. Two major milestones in the initial development of positional cloning are usually referred: in 1983 the localization on chromosome 4 of the gene responsible for Huntington disease, a late onset polyglutamine neurodegenerative disorder [6] and the identification, six years later, of the causative gene of cystic fibrosis, a channelopathy inherited as autosomal recessive [7, 8].

Although potentially successful in its application to single-gene diseases, linkage analysis displays a reduced potential for identifying genes involved in complex traits [9]. The difficulties encountered in the analysis of multifactorial phenotypes justified the development and implementation of association studies, which can roughly be divided as those: a) performed at the population level, with allele frequencies of a particular locus being compared between cases and unrelated control; and b) based on families, where samples from the index-case and the respective parents and siblings are necessary in order to compare the frequencies of alleles transmitted to affected and non-affected children. Association studies may also use a candidate gene approach, looking for the relationship between genetic variants previously selected, and the phenotype under consideration. Candidate or direct studies are focused on genes that are selected given “a priori” assumptions about a functional or causative role in disease. Criticisms to this approach have been based on the fact that there is still an incomplete understanding of pathogenic processes to allow a proper candidate selection. Research conducted in genetically isolated populations has the potential to increase the rate of success of gene-finding studies, allowing researchers to circumvent some of the previously reported constraints.

5.2. HUMAN GENETIC ISOLATES: GENERAL CHARACTERISTICS

Genetically isolated populations, also referred in the literature as “population isolates” or “founder populations” [10] have derived from a reduced number of individuals, originally belonging to a main parental population [1, 11]. There is no consensus definition of “genetic isolate”; criteria used for defining an isolate are focused mainly on the number of initial founders and on the existence of some extent of isolation. A non-exhaustive list of the main human population isolates is provided in Table **1**. Some populations that do not strictly follow all aspects of a “classic” isolate are also included in Table **1**; such populations present a reduction of variability, as compared to other European populations, and are sometimes named “non-extreme genetic isolates” or “non-conventional isolates” [12]. The distinctiveness of some non-conventional isolates has been questioned: Garagnani and collaborators [13], for example, have refuted the “uniqueness” of the Basques. Despite the lack of consensus, several studies support that in such non-conventional isolated populations the occurrence of genetic drift, with the consequent loss of diversity, has had a higher impact that in other European populations; the study of Helgason and colleagues [12], for example, is in agreement with such point of view on what concerns the Icelanders. To define genetic isolates, the determination of the homozigosity/heterozigosity patterns exhibited is considered of value, with highly inbred populations presenting increased frequencies of homozygosity.

Table 1. Non-exhaustive list of genetically isolated populations. Examples of non-conventional isolated populations are also provided.

Population	Approximate Age (Years)	Reference
Genetic Isolates		
Finland	2000*	[14]
Sardinia	Pre-Neolithic	[15]
Quebec (French Canadian)	400	[16]
Hutterite	145	[17]
Mennonite	300	[18]
Amish	290	[18]
Paisa Community (Colombia)	365	[19]
Pima Indians	2000	[20]
Ashkenazi Jews	500	[21]
Tristan da Cunha	200	[22]
Non-conventional isolates		
Iceland	1100	[23]
Newfoundland	300	[24]
Basque Country	7000	[25]

*early settlement

Several types of founding events can be on the basis of the constitution of genetic isolates: the settlement of a new area, mortality due to epidemics or even the existence of social or religious barriers are some examples [11]. Regardless of the originating cause, founder populations need to be isolated for several generations, with limited genetic interchange with other subpopulations [26]. Isolation plays an important role in shaping the population genetic constitution, to an extent which is modulated by distinct factors, namely the isolate age, the total number of founders, the number and intensity of bottlenecks, as well as the amount of endogamy. Isolates tend to display higher levels of endogamy and restricted gene flow with surrounding populations; as a consequence, they show higher levels of genetic homogeneity, when compared to cosmopolitan populations [11]. Noteworthy, as previously referred, the high frequency of inbreeding seldom observed in isolated populations tends to produce an increased incidence of recessive disorders [27].

Due to a combination of mechanisms, otherwise rare genetic variants can reach in isolates considerable frequencies. Palotie *et al.* [28] have demonstrated that the chances of detecting rare variants in isolated populations is increased; these authors estimated that such rare alleles can achieve, in population isolates, frequencies up to 5%. An example is the *TBC1D4* (TBC1 Domain Family, Member 4) gene, which has been associated with diabetes mellitus, and whose p.Arg684Ter nonsense variant is frequent in Greenland (17%), but almost inexistent in other global populations [29]. Noteworthy, whereas certain alleles increase their frequency others can, on the contrary, reach extinction, thus contributing to the reduction of genetic variability, a hallmark of genetic isolates.

Multigenerational pedigrees (seldom called “mega pedigrees”), frequently seen in population isolates, can display multiple cases of patients with rare diseases, which share a homogeneous phenotype, under similar environmental conditions. The framing of computerized genealogical databases in isolated populations provides an important tool based on which several types of studies can be performed; one example of a population that has been studied using extensive genealogical information is the French-Canadian Quebec, whose genealogical records have been massively computerized, tracing back any native of Quebec to the founding settlers [30]. Quebec has, furthermore, been the target of wider historical and anthropological studies, which have provided a conceptual basis for the interpretation of the genetic data. Another example of a population whose genealogical records have been extensively analyzed is the Iceland population;

this non-extreme genetic isolate has been the target of the private company DeCODE Genetics, which has built a database currently including over half of the total adult population. Encrypted information of all participants in the Iceland database can be cross-referenced with genealogical information, similarly encrypted [31]. Due to the availability of such tools, in a study of type 2 diabetes, performed in Iceland, extended genealogical information allowed inference of genotype probabilities in an extended number of untyped individuals [32].

Until the emergence of the major human genetic diversity projects (for a review on the main human genome projects see chapter 1), the knowledge concerning the genetic composition of human populations was limited. The HapMap Project, one of the main genome-scale initiatives, has made available a map of common sequence variants [33]. The data produced by HapMap concerning the haploblock structure of the human genome has had a great impact in human genetic studies, namely on gene-finding efforts. Another prominent genome initiative is the “1000 genome project” [34]. Results from this project have confirmed that linkage disequilibrium (LD) spans over a larger extension in population isolates, compared with non-isolated populations [11]. Because LD intervals are conditioned by the amount of recombination, which in turn is modulated by the age of the population, younger isolates in particular are expected to display larger LD blocks [4], having therefore an increased potential for gene finding studies.

5.3. THE USE OF POPULATION ISOLATES IN GENE FINDING STUDIES: ADVANTAGES AND LIMITATIONS

The most powerful methodologies for identifying genes related with complex diseases require a large number of samples [35], a demand which can be difficult to meet in several situations. Given such constraints, and in view of the success obtained for gene finding in Mendelian disorders, the scientific community was encouraged to use isolated populations with the purpose of complex disease analysis; in alternative to large samples sizes, these populations should provide an appropriate study design, in combination with rigorous statistical methods [28, 36].

Human genetic isolates display several recognized advantages for gene finding studies:

a. Environmental and cultural homogeneity: individuals belonging to isolated populations tend to share lifestyle habits, such as diet and physical activity and are exposed to similar environmental conditions; this homogeneity is considered as extremely positive, since it improves the signal-to-noise ratio [4];
b. Reduced phenotypic heterogeneity: patients share ancestors and should display more homogeneous clinical presentations. Phenotype definition and diagnosis harmonization are, furthermore, considered to be facilitated [37];
c. Particularly in the case of rare diseases, a sufficient number of cases can be frequently only found in isolates; for such patients extended well documented genealogical information is usually available and can be cross-referenced to clinical data;
d. High participation rate in studies: the restricted geographical distribution enhances the possibility of successful follow ups to collect samples [26];
e. Higher levels of LD: because longer LD blocks imply longer haplotypes, association studies are facilitated and fewer markers are necessary. The age of the isolate is an important factor, as previously referred, since LD intervals can reach up to 1Mb in alleles of young isolates [37];
f. The enrichment of certain alleles by a combination of mechanisms empowers the discovery of genetic variants using more reduced discovery sets [38]. Although allelic heterogeneity can also exist associated with common diseases even in isolated populations, it is expectable that a reduced number of predisposing alleles should exist in isolates. The amplification of certain genetic variants, as compared to the parental population, whilst keeping a homogeneous genetic background [10], increases power of genetic association studies [39].

Successful applications of isolates in gene identification have used several strategies, whose description is outside the scope of this chapter. Noteworthy, one of the successful strategies is homozygosity mapping; this approach profits from the high frequency of consanguineous unions in population isolates, postulating that alleles present in homozygous patients for recessive disorders are likely to be derived from a common ancestor [2], causing markers close to the gene of interest to also display homozygosis. Finding disease causing genes in this situation implies finding shared regions between all patients, which should show homozygosity for markers. Because in outbreed populations genetic heterogeneity is likely to complicate this scenario, the approach is particularly powerfully in genetic isolates [2].

Although the importance of genetic isolates for gene finding studies is well established, drawbacks have also to be seriously considered:

a. Several genetic isolates have been described, belonging to various distinct backgrounds and with different demographic histories. For some isolates reliable information on the initial genetic constitution is lacking; thus the number of founders, the extent as well as the duration of the isolation might be not very well known [37]. The strategy to use in gene finding efforts has therefore to be in line with what is known about the population. As a general rule, the better the characteristics relative to the population are established the easier it will be to define an ideal strategy for disease identification [37];
b. One important concern is that findings derived from isolated populations might not be generalized to other populations [39]. This lack of reproducibility is acknowledged as a major problem in association studies [40]. Zeggini argued that even in cases where findings cannot be replicated, the associations detected should allow further understanding of the biology of the disease [39];
c. Strong LD might difficult the distinction of the biologically significant variants from variants with no effect in complete LD with them [1]. Although the value of population isolates for the initial stages of gene finding studies – finding a locus with a limited number of markers - is consensual and predicted to be based on extended LD, fine mapping (fine relevant mutation identification) can be hampered by the difficulties in distinguishing biologically relevant variables from non-significant variants in LD with them. Gene finding studies would therefore benefit from the analysis in multiple populations with distinct demographic structure. One strategy would be to initiate the genome-wide mapping stage in population sub-isolates, and then proceed for the fine mapping stage using the main population to which sub-isolates belong to [37]. Fine-mapping of the causal variants has been proposed as being possible thru transethnic meta-analysis [11];

d. The results from several studies which estimated the extent of LD in isolated populations have been analyzed in several papers, namely in [4]. The fact that in many of this studies no significant differences in the length of LD has been detected, as compared to more opened, mixed populations, could be interpreted as a result of the poor definition of the samples analyzed as representative of such isolates; implications, however, have to be carefully considered.

5.4. THE GENOMIC ERA AND THE USE OF ISOLATED POPULATIONS FOR GENE FINDING

The advent of high-throughput sequencing and the development of the necessary computational tools had a profound effect in gene-finding efforts; whole exome and whole genome sequencing studies (WES and WGS, respectively) have been allowing, in the recent years, the discovery of several genes associated with monogenic disorders. The impact of genomics on advances concerning Mendelian conditions is estimated to be tremendous, and initiatives such as the Centers for Mendelian Genomics will accelerate discoveries based on powerful large scale data [41]. The use of population isolates for gene finding studies has been maximized with such genomic tools, since WGS based studies conducted in isolated populations are feasible with a smaller set of samples, without jeopardizing discovery power [39, 42]. Gene finding studies using genome sequencing in members of extended families from isolates will look for similar chromosomal segments among relatives from whom risk variants can be identified [43].

In a first stage, genomic data from the isolated population can provide in-depth insights into its structure, potentially allowing research designs which are particularly suited to each isolate. Although genome-scale analysis only recently has entered a more widespread stage, the acknowledgment of the importance of large-scale comparison of DNA sequences in distinct populations – the so called "population genomics" - and the impact that it should have in several areas, namely in gene finding efforts, has been highlighted by Jorde and collaborators in 2001 [44]. Drawing the attention to the fragilities of isolates, Jorde and co-authors [44] further highlighted that to circumvent constraints such as reduced sample size and lower levels of markers heterozygosity (usually found in isolated populations), genomic studies should provide the necessary information on the genetic makeup of the isolates. With the accessibility of genome scale studies, this has proven to be an important point: in their study of the Iceland population, for example, Gudbjartsson and colleagues sequenced the whole genome of over 2500 Icelanders, demonstrating an excess of homozygosity and rare protein-coding variants in this population [45].

Important studies at the genome scale using genetic isolates are already available. For example, in an investigation undertaken with the aim to dissect the genetic basis of Parkinson disease (PD) Quadri and collaborators [46] performed exome sequencing in 100 unrelated patients and in a large Sardinian control sample. After performing validation by Sanger sequencing and genotyping patients from 3 large cohorts of distinct backgrounds, the authors identified novel moderately rare variants in several genes, which were either specifically present in PD patients or reached higher frequencies in the group of patients, thus considering them as novel candidate risk genes for PD.

CONCLUDING REMARKS

Human genetic isolates provide an exclusive opportunity to unveil the genetic basis of disease. As high-throughput sequencing allows an in-depth investigation of the genetic makeup of these isolates their potential for gene discovery is boosted, since the information produced can be used to fine-tune the strategies to be used. As the cost with WGS drops and more sophisticated quality control, data management and bioinformatics tools emerge, the value of genetic isolates continues to increase.

CONFLICT OF INTEREST

The author confirms that author has no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Kristiansson K, Naukkarinen J, Peltonen L. Isolated populations and complex disease gene identification. Genome Biol 2008; 9(8): 109.
[http://dx.doi.org/10.1186/gb-2008-9-8-109] [PMID: 18771588]

[2] Sheffield VC, Stone EM, Carmi R. Use of isolated inbred human populations for identification of disease genes. Trends Genet 1998; 14(10): 391-6.
[http://dx.doi.org/10.1016/S0168-9525(98)01556-X] [PMID: 9820027]

[3] Online Mendelian Inheritance in Man, OMIM® [homepage on the Internet]. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. Available from: http://omim.org/ 2016.

[4] Heutink P, Oostra BA. Gene finding in genetically isolated populations. Hum Mol Genet 2002; 11(20): 2507-15.
[http://dx.doi.org/10.1093/hmg/11.20.2507] [PMID: 12351587]

[5] Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 2002; 3(5): 391-7.
[http://dx.doi.org/10.1038/nrg796] [PMID: 11988764]

[6] Gusella JF, Wexler N, Conneały P, *et al.* A polymorphic DNA marker genetically linked to HD. Nature 1983; 306(5940): 234-8.
[http://dx.doi.org/10.1038/306234a0] [PMID: 6316146]

[7] Riordan JR, Rommens JM, Kerem B, *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 1989; 245(4922): 1066-73.
[http://dx.doi.org/10.1126/science.2475911] [PMID: 2475911]

[8] Kerem B, Rommens JM, Buchanan JA, *et al.* Identification of the cystic fibrosis gene: genetic analysis. Science 1989; 245(4922): 1073-80.
[http://dx.doi.org/10.1126/science.2570460] [PMID: 2570460]

[9] Mayeux R. Mapping the new frontier: complex genetic disorders. J Clin Invest 2005; 115(6): 1404-7.
[http://dx.doi.org/10.1172/JCI25421] [PMID: 15931374]

[10] Lewis R. Founder populations fuel gene discovery. Scientist 2001; 15(8): 8.

[11] Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. Brief Funct Genomics 2014; 13(5): 371-7.
[http://dx.doi.org/10.1093/bfgp/elu022] [PMID: 25009120]

[12] Helgason A, Nicholson G, Stefánsson K, Donnelly P. A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. Ann Hum Genet 2003; 67(Pt 4): 281-97.
[http://dx.doi.org/10.1046/j.1469-1809.2003.00046.x] [PMID: 12914564]

[13] Garagnani P, Laayouni H, González-Neira A, *et al.* Isolated populations as treasure troves in genetic epidemiology: the case of the Basques. Eur J Hum Genet 2009; 17(11): 1490-4.
[http://dx.doi.org/10.1038/ejhg.2009.69] [PMID: 19417765]

[14] Polvi A, Linturi H, Varilo T, *et al.* The Finnish disease heritage database (FinDis) update-a database

for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. Hum Mutat 2013; 34(11): 1458-66.
[http://dx.doi.org/10.1002/humu.22389] [PMID: 23904198]

[15] Angius A, Melis PM, Morelli L, *et al.* Archival, demographic and genetic studies define a sardinian sub-isolate as a suitable model for mapping complex traits. Hum Genet 2001; 109(2): 198-209.
[http://dx.doi.org/10.1007/s004390100557] [PMID: 11511926]

[16] Scriver CR. Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet 2001; 2: 69-101.
[http://dx.doi.org/10.1146/annurev.genom.2.1.69] [PMID: 11701644]

[17] Pichler I, Fuchsberger C, Platzer C, *et al.* Drawing the history of the hutterite population on a genetic landscape: inference from Y-chromosome and mtDNA genotypes. Eur J Hum Genet 2010; 18(4): 463-70.
[http://dx.doi.org/10.1038/ejhg.2009.172] [PMID: 19844259]

[18] Strauss KA, Puffenberger EG. Genetics, medicine, and the plain people. Annu Rev Genomics Hum Genet 2009; 10: 513-36.
[http://dx.doi.org/10.1146/annurev-genom-082908-150040] [PMID: 19630565]

[19] Carvajal-Carmona LG, Soto ID, Pineda N, *et al.* Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am J Hum Genet 2000; 67(5): 1287-95.
[http://dx.doi.org/10.1086/321216] [PMID: 11032790]

[20] Thompson DB, Ravussin E, Bennett PH, Bogardus C. Structure and sequence variation at the human leptin receptor gene in lean and obese Pima Indians. Hum Mol Genet 1997; 6(5): 675-9.
[http://dx.doi.org/10.1093/hmg/6.5.675] [PMID: 9158141]

[21] Ostrer H. A genetic profile of contemporary Jewish populations. Nat Rev Genet 2001; 2(11): 891-8.
[http://dx.doi.org/10.1038/35098506] [PMID: 11715044]

[22] Soodyall H, Nebel A, Morar B, Jenkins T. Genealogy and genes: tracing the founding fathers of Tristan da Cunha. Eur J Hum Genet 2003; 11(9): 705-9.
[http://dx.doi.org/10.1038/sj.ejhg.5201022] [PMID: 12939658]

[23] Smith KP. Landnmám: the settlement of Iceland in archaeological and historical perspective. World Archaeol 1995; 26(3): 319-47.
[http://dx.doi.org/10.1080/00438243.1995.9980280]

[24] Rahman P, Jones A, Curtis J, *et al.* The Newfoundland population: a unique resource for genetic investigation of complex diseases. Hum Mol Genet 2003; 12(Spec No 2): R167-72.
[http://dx.doi.org/10.1093/hmg/ddg257] [PMID: 12915452]

[25] Behar DM, Harmant C, Manry J, *et al.* The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. Am J Hum Genet 2012; 90(3): 486-93.
[http://dx.doi.org/10.1016/j.ajhg.2012.01.002] [PMID: 22365151]

[26] Arcos-Burgos M, Muenke M. Genetics of population isolates. Clin Genet 2002; 61(4): 233-47.
[http://dx.doi.org/10.1034/j.1399-0004.2002.610401.x] [PMID: 12030885]

[27] Peltonen L. Positional cloning of disease genes: advantages of genetic isolates. Hum Hered 2000; 50(1): 66-75.
[http://dx.doi.org/10.1159/000022892] [PMID: 10545759]

[28] Palotie A, Widén E, Ripatti S. From genetic discovery to future personalized health research. N Biotechnol 2013; 30(3): 291-5.
[http://dx.doi.org/10.1016/j.nbt.2012.11.013] [PMID: 23165095]

[29] Moltke I, Grarup N, Jørgensen ME, *et al.* A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. Nature 2014; 512(7513): 190-3.
[http://dx.doi.org/10.1038/nature13425] [PMID: 25043022]

[30] Balsac Population Database [homepage on the Internet]. Québec: Projet BALSAC. Available from: http://balsac.uqac.ca/ 2016.

[31] Hakonarson H, Gulcher JR, Stefansson K. deCODE genetics, Inc. Pharmacogenomics 2003; 4(2): 209-15.
[http://dx.doi.org/10.1517/phgs.4.2.209.22627] [PMID: 12605555]

[32] Steinthorsdottir V, Thorleifsson G, Sulem P, *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 2014; 46(3): 294-8.
[http://dx.doi.org/10.1038/ng.2882] [PMID: 24464100]

[33] The International HapMap Project. Nature 2003; 426(6968): 789-96.
[http://dx.doi.org/10.1038/nature02168] [PMID: 14685227]

[34] Abecasis GR, Altshuler D. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319): 1061-73.

[35] Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annu Rev Genet 2010; 44: 293-308.
[http://dx.doi.org/10.1146/annurev-genet-102209-163421] [PMID: 21047260]

[36] Hou L, Faraci G, Chen DT, *et al.* Amish revisited: next-generation sequencing studies of psychiatric disorders among the Plain people. Trends Genet 2013; 29(7): 412-8.
[http://dx.doi.org/10.1016/j.tig.2013.01.007] [PMID: 23422049]

[37] Varilo T, Peltonen L. Isolates and their potential use in complex gene mapping efforts. Curr Opin Genet Dev 2004; 14(3): 316-23.
[http://dx.doi.org/10.1016/j.gde.2004.04.008] [PMID: 15172676]

[38] Service S, DeYoung J, Karayiorgou M, *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 2006; 38(5): 556-60.
[http://dx.doi.org/10.1038/ng1770] [PMID: 16582909]

[39] Zeggini E. Using genetically isolated populations to understand the genomic basis of disease. Genome Med 2014; 6(10): 83.
[http://dx.doi.org/10.1186/s13073-014-0083-5] [PMID: 25473423]

[40] Chanock SJ, Manolio T, Boehnke M, *et al.* Replicating genotype-phenotype associations. Nature 2007; 447(7145): 655-60.
[http://dx.doi.org/10.1038/447655a] [PMID: 17554299]

[41] Bamshad MJ, Shendure JA, Valle D, *et al.* The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A 2012; 158A(7): 1523-5.
[http://dx.doi.org/10.1002/ajmg.a.35470] [PMID: 22628075]

[42] Gilly A, Ritchie GR, Southam L, *et al.* Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. Hum Mol Genet In press
[http://dx.doi.org/10.1093/hmg/ddw088] [PMID: 27146844]

[43] Hou L, Faraci G, Chen DT, *et al.* Amish revisited: next-generation sequencing studies of psychiatric disorders among the plain people. Trends Genet 2013; 29(7): 412-8.
[http://dx.doi.org/10.1016/j.tig.2013.01.007] [PMID: 23422049]

[44] Jorde LB, Watkins WS, Bamshad MJ. Population genomics: a bridge from evolutionary history to genetic medicine. Hum Mol Genet 2001; 10(20): 2199-207.
[http://dx.doi.org/10.1093/hmg/10.20.2199] [PMID: 11673402]

[45] Gudbjartsson DF, Helgason H, Gudjonsson SA, *et al.* Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 2015; 47(5): 435-44.
[http://dx.doi.org/10.1038/ng.3247] [PMID: 25807286]

[46] Quadri M, Yang X, Cossu G, *et al.* An exome study of Parkinson's disease in sardinia, a Mediterranean genetic isolate. Neurogenetics 2015; 16(1): 55-64.
[http://dx.doi.org/10.1007/s10048-014-0425-x] [PMID: 25294124]

CHAPTER 6

Complex Human Phenotypes: The Interplay between Genes and Environment

Mar Fatjó-Vilas[1,2,3,4] and **Bárbara Arias**[2,3,4,*]

[1] *FIDMAG Germanes Hospitalàries Research Foundation. Sant Boi de Llobregat, Barcelona, Spain*

[2] *Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain*

[3] *Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain*

[4] *Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain*

Abstract: Genetic epidemiology is the discipline that studies the role of genetic and environmental factors in the origin of complex traits, behaviors and diseases. The major focus of genetic epidemiology is the analysis of the relative contribution of genes, environment and their interplay in human traits. Among others, twin studies have become an important tool to disentangle the different roles of genes and environment and to estimate heritability. Within the genetic epidemiology, the ecogenetics study the relationship between genetic and environmental factors and seek to understand both the vulnerability of different genotypes present in the population facing the same environmental risk factors (gene-environment interaction, GxE) and the influence of the individual's own genotype in the search of specific environments and/or risk factors (gene-environment correlation, rGE). There are numerous studies from quantitative genetics and molecular genetics that describe such GxE and rGE effects on the etiology of complex traits and disorders. However, it is important to consider the methodological requirements and limitations associated with these studies. Undoubtedly one of the challenges of genetic epidemiology in the coming years will be to combine the gene-environment studies (based on specific assumptions) with the huge amount of genomic data provided by new molecular approaches.

[*] **Corresponding author Bárbara Arias:** Unitat d'Antropologia, Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals. Facultat de Biologia, Universitat de Barcelona. Av. Diagonal 643, 08028 Barcelona, Spain; Tel: +34934021460; E-mail: barbara.arias@ub.edu

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

Keywords: Antisocial behaviour, Cannabis use, Childhood maltreatment, Complex traits, Environment, Friendship, Genes, Gene-environment correlation, Gene-environment interaction, Heritability, Obesity, Schizophrenia, Twins.

6.1. INTRODUCTION

Nowadays, there is consensus that all traits of human development are influenced by both genetic and environmental factors; which represents a change from previous perspectives [1]. At 20th century, genes and environment were believed to play independent roles, being the interaction an exception rather than the rule [2]. However, it has been noted that genes and environments interact and influence each other to shape human development (as in any other living organism) [3, 4]. Since then, evidence has accumulated showing that genes and environment work co-jointly rather than independently on human development.

Accordingly, those traits in which multiple genetic and environmental factors are implicated, are known as *complex traits*. This term is used to define any trait that does not show a classic Mendelian recessive, dominant or X-linked inheritance attributable to a single gene locus. Then, the architecture of a complex trait comprises all the genetic and environmental factors that affect the trait, along with the interactions among the factors.

Therefore, complex traits are not manifested in a few easily distinguished categories. Instead, they vary continuously from one phenotypic extreme to the other, with no clear-cut breaks in between. Human height, blood pressure or intelligence quotient can be considered such kind of traits. Such traits are called continuous because there is a continuous gradation from one phenotype to the next.

Continuous traits were first investigated by biometricians leaded by Sir Francis Galton, a cousin of Charles Darwin, by 1900. When Mendel’s work was rediscovered, Biometricians pointed out that most of the characters likely to be important in evolution (such as body size, build, strength or finding food) were continuous or quantitative characters and, then, they were not explained by Mendelian inheritance. The controversy run on between the two positions until 1918, when a seminal paper by R.A. Fisher demonstrated that a complex trait can be explained by Mendelian inheritance if a large number of independent genetic factors (polygenic characters) affect the trait. In that case, the trait would exhibit the continuous nature, quantitative variation and family correlation previously described by the biometricians [5].

The task of disentangling genetic from environment impacts has proved extremely difficult [6]. Quantitative genetic methods, such as family, twin and adoption studies have firmly contributed to establish the implication of genes and environment in several human traits and diseases. In this chapter, we will comment the special relevance of twin studies. Furthermore, to understand the relationship between genetic and environmental factors, both from quantitative and molecular data, in this chapter we will discuss gene-environment correlation and interaction mechanisms.

6.2. TWIN STUDIES: DISENTANGLING THE INVOLVEMENT OF GENES AND ENVIRONMENT IN COMPLEX TRAITS

6.2.1. Twinning in Human Populations

There are two types of twins: monozygotic (MZ) or identical and dizygotic (DZ) or fraternal twins. MZ twins result from a single fertilized egg that splits for unknown reasons, producing two (or sometimes more) genetically identical individuals. For this reason, MZ twins are said to be natural clones. They are supposed to be 100% genetically identical at the DNA sequence level [7, 8]. DZ or fraternal twins occur when two eggs are separately fertilized; they have different chorions and amnions. Like other siblings, they are on average 50% similar genetically.

As a curiosity, in human pregnancies as many as 20 percent of foetuses are twins, but because of complications associated with these pregnancies, often one member of the pair does not progress very early in pregnancy [8]. The prevalence range of spontaneous twinning in live births varies worldwide: the lower frequency corresponds to Asia (about 6 in 1000); Europe and the USA show an intermediate prevalence (about 10-20 in 1000) and the higher rates are found in Africa (about 40 in 1000). More specifically, in Japan only 1 in 250 newborn babies is a twin, whereas in Nigeria 1 in 11 is a twin [9].

In Spain, the twinning rate in 1980 was 7.4 per 1000 deliveries. Thereafter, a continuous increase in the number of multiple deliveries occurred, increasing to 12.4 in 1996 [10]. This rate increase could be related to the change in age structure of mothers [11], as well as to assisted reproduction techniques [10]. Other factors apart from the mother's age could be influencing the twinning rate such as birth order, season, socioeconomic factors, ethnic group, and/or rural–urban differences [10].

6.2.2. Twin Studies

Apart from adoption studies based on singletons (genetically related individuals such as siblings who do not share a common family environment), twin studies constitute one of the most powerful methods used to disentangle genetic from environmental sources of resemblance between relatives [7, 8, 12]. Nowadays, large twin registers are established in different countries around the world which collect a wide range of traits, environmental and biological data in twins as well as their family members' [7, 12].

Classical twin studies compare phenotypic similarities (or concordance) of MZ and DZ twins [7, 13]. By comparing the similarity/concordance between MZ twins for a trait or disease with those of DZ twins it is possible to estimate to what extent genetic and phenotypic variation are inter-related. If genetic factors are important for a trait, identical twins must be more similar than fraternal twins.

In the classical twin design, one can infer the relative contribution of genetic and environmental factors by comparing the observed correlations (or concordance) between twin members (Fig. **1**).

The sources usually estimated of genetic and environmental variation in behaviour genetics are: *additive genetic influences (A)*, *shared or common environmental influences (C)* and *non-shared or unique environmental influences (E)*. *A* represents the sum of the effect of the individual alleles at all loci that influence a trait. *A* is also known as *heritability* (h^2); this concept is further discussed below. *C* includes environmental influences that contribute to similarity within twin pairs, while *E* represents environmental influences that are unique to each individual, plus measurement error (Fig. **2**) [8, 13].

A

MZ and DZ Correlations						
Trait	r_{MZ}	r_{DZ}		***Trait***	r_{MZ}	r_{DZ}
Attention-deficit	0.83	0.42		Memory	0.45	0.34
Blood pressure	0.81	0.44		Peptic ulcer	0.65	0.35
Bone mineral density	0.90	0.58		Plasma cholesterol	0.80	0.68
Cerebral hemisphere volume	0.87	0.56		Premenstrual syndrome	0.55	0.28
Extraversion	0.51	0.26		Spatial ability	0.65	0.45
General intelligence	0.86	0.60		Spatial reasoning	0.65	0.45
Height	0.98	0.57		Verbal ability	0.76	0.43

B

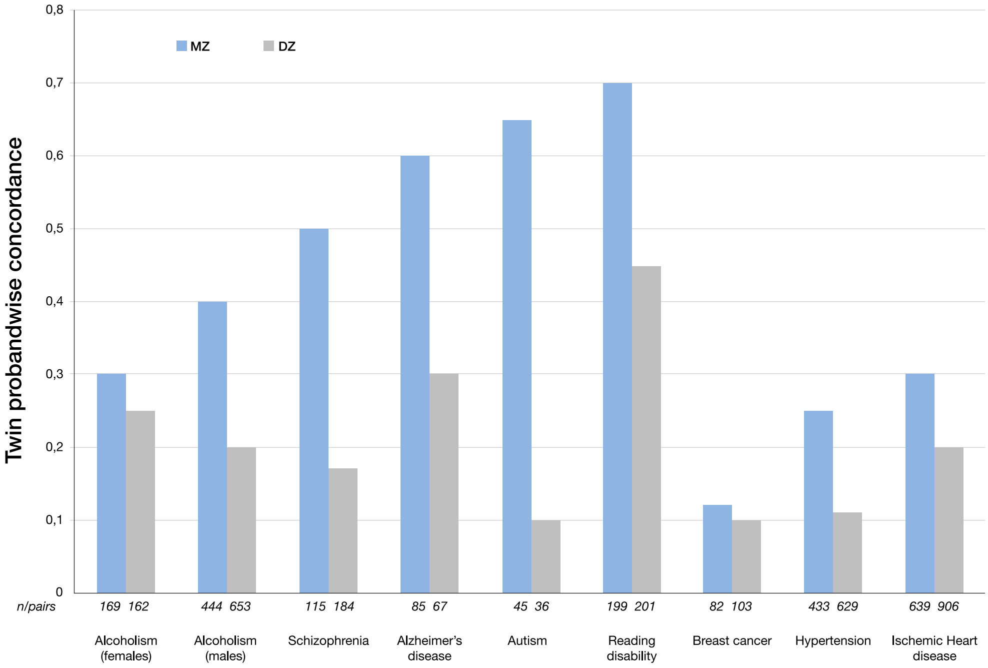
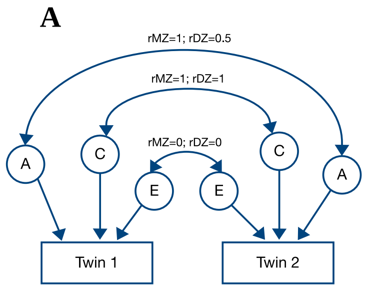


Fig. (1). According to the classical twin design, correlations (for quantitative traits) and concordances (qualitative traits) within twin pairs are estimated in order to infer the relative contribution of genetic and environmental factors to a complex trait variance. (**A**) The table shows the correlation (r) of several quantitative traits in monozygotic (MZ) twin pairs and dizygotic (DZ) twin pairs. (**B**) The graph shows the MZ and DZ twins concordances for several behavior disorders and medical conditions. Adapted from Plomin *et al.* 1994 [14].

According to this model, known as ACE model, the total phenotypic variance (P) of a given trait is the sum of A, C and E variance components (P=A+C+E). As MZ and DZ twins show different degrees of correlation for the genetic component (A) but similar degrees of correlation for the environmental components (C and E), twin data allows estimating the different variance components. MZ twin pairs correlate 1 for A (they are assumed to be genetically identical), whereas DZ twin pairs correlate 0.5. Both MZ and DZ pairs correlate 1 for C and 0 for E (because C contributes to make them more alike and E to make them more different) (Fig. **2**).



B

Trait	*a*^{2}	*c*^{2}	*e*^{2}
Aggression	0.44	0.06	0.50
Cerebral Volume	0.94	0.06	
Heart rate (beats/minute)	0.23	0.23	0.54
Schizophrenia	0.88		0.12

Fig. (2). (**A**) Path diagram for the basic univariate twin model. Circles represent latent variables (A, C and E), and rectangles represent the observed and measured traits (phenotypes). In the diagram, the phenotype of each twin is decomposed into A, C and E variance components corresponding to additive genetic, shared environment and non-shared environmental influences respectively. Additive genetic factors correlate 1 for MZ and 0.5 for DZ pairs. Shared environmental factors correlate 1 for both MZ and DZ pairs and non-shared environmental factors are uncorrelated (r=0). (**B**) Table showing some examples of ACE model in complex traits: a^2 (additive genetic), c^2 (common environmental) and e^2 (individual-specific environment).

The validity of the classical twin method depends on several assumptions. One of them is the equal environment assumption (EEA), that is, the assumption that MZ and DZ are equally correlated in their exposure to environmental factors of etiological importance for the trait that is being studied [15]. A possible violation of the EEA can arise if MZ twins are treated more similarly than DZ twins and because of that, the resemblance within MZ increases. This violation would result in an over-estimation of the heritability of the observed trait. However, this assumption is continuously tested and different studies have found support for its validation in a variety of behavioural traits and disorders [16, 17].

6.2.3. The Heritability Concept

Beyond determining whether there is a genetic influence on a trait or not, it is important to "quantify the role of genes". It refers to which proportion of the phenotypic differences for a particular trait observed in a population at a given moment can be attributed to genetic differences between individuals. And this fraction of phenotypic variance explained by genetic differences between individuals is called heritability (h^2) [18].

In the classical method based on twins, Falconer's formula was used to estimate heritability (*i.e.* A) based on twin correlations: $h^2 = 2(r_{MZ} - r_{DZ})$ or based in twin concordances: $h^2 = 2(C_{MZ} - C_{DZ})$ [13]. The value of h^2 ranges from 0 when the genes have no influence, and 1, when genes are fully responsible for the phenotypic variance (Table **1**). Twin statistical analysis methods are constantly evolving and improving, and the use of this formula has been replaced by structural equation models that also allow including more than one trait and thus analysing the genetic and environmental factors shared between different features.

Table 1. Broad-sense heritability values, in percent, based on twin studies.

Trait	Heritability (h^2)	Trait	Heritability (h^2)
Longevity	29	Verbal ability	63
Height	85	Numerical ability	76
Weight	63	Memory	47
Amino acid excretion	72	Sociability index	66
Serum lipid levels	44	Temperament index	58
Maximum blood lactate	34	Maximum heart rate	84

A high h^2 implies strong correlation between phenotype-genotype thus loci with an effect on the trait can hypothetically be more easily detected. However, it is important to highlight that h^2 by itself does not provide information on the genetic architecture of the features; *i.e.* do not tell us what or how many genes are responsible for them. Moreover, h^2 do not consider additive effects of the genetic and environmental factors, and, hence, the estimate does not consider interaction (gene-gene, gene-environment or environment-environment) which are likely to be common in the complex traits or phenotypes. Then, although these and other limitations are acknowledged by researchers; the limitations do not negate the usefulness of twin studies and their contribution in the understanding the heritability of complex traits.

6.3. GENE-ENVIRONMENT INTERPLAY MECHANISMS

6.3.1. Gene-Environment Interaction

That genetic and environmental factors are interrelated and jointly modulate the expression of multiple phenotypes is a completely expected and plausible biological mechanism; *i.e.* genotypes do not exist "in a vacuum" and their expression must depend to some extent on environmental context (from the cell-environment to the social-environment of a person) [19].

Whereas the general term "gene-environment interplay" refers to a variety of concepts (Rutter, Moffitt, & Caspi, 2006), the two most frequently studied genetic-environmental mechanisms are gene-environment interactions (GxE) and gene–environment correlations (rGE).

There are many ways of thinking about GxE [20], but in quantitative genetics the term generally means that genetic differences between individuals explain part of the variance in environmental responses [8, 21 - 23]. In other words, GxE refers to the genetic control of sensitivity to the environment (Fig. **3**) [24]. In the simplest cases, environmental effects on a phenotype are additive, and each environment adds (or detracts) the same amount to the phenotype independently of the genotype. When this is not true, the environmental effects on the phenotype differ according to genotype, and genotype-environment interaction is present. In some cases, GxE can even change the relative ranking of the genotypes, so a genotype that is superior in one environment may become inferior in another [25].

In biological terms, GxE can be defined as the joint effect of one or more genes with one or more environmental factors that cannot be readily explained by their separate effects [27]. Interaction between genes and environment means more than simply stating that both are involved in a trait or disease aetiology [24]. This is also related to the terms synergism and parallelism. Biological synergism refers to the proportion of the population exposed to both genes and environment that developed a specific complex trait or behaviour specifically because of the combination of these exposures; while parallelism refers to the proportion of the population exposed to both genes and environment that developed a specific complex trait because of either genes or environment [28].

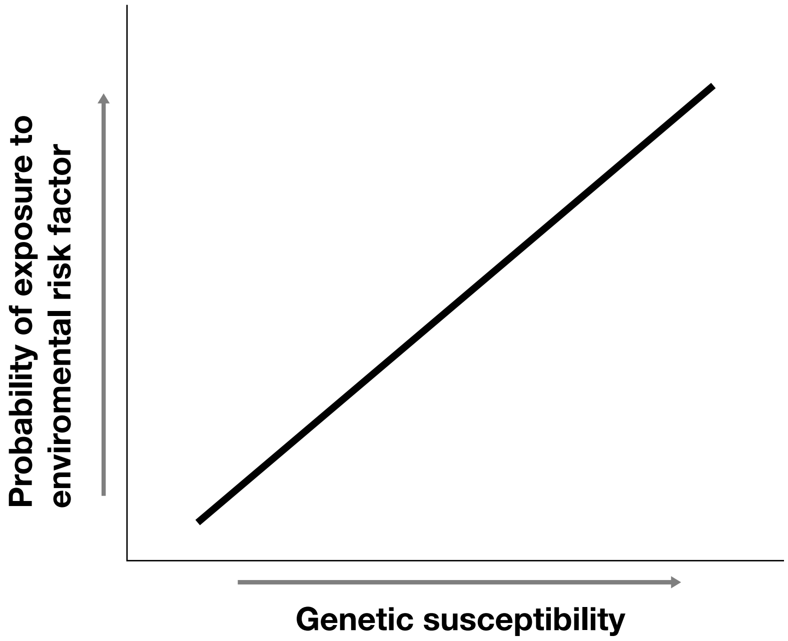


Fig. (3). Gene-environment interaction: the genetic control of sensitivity to the environment. The effect of an environmental factor on the risk for a disease (or on the expression of a quantitative trait) is modulated by the genotype. Then, with the same exposition to an environmental factor, one individual with the genotype + would have higher risk than one carrying the genotype -. Adapted from van Os and Marcelis 1998 [26].

Within epidemiology, the term interaction is used to define the situation where the relationship between one exposure (risk factor) and a disease is dependent on the presence or absence of another exposure [29].

Also, some authors have pointed out that interaction is model dependent, meaning that the model used to test the interaction affects the interpretation of the results. In this regard, interaction effects can be modelled on either *additive* or *multiplicative* scales [29]. Additive models use risk differences (RD) while multiplicative models use risk ratios (RR). For example, an additive model of the statistical interaction between A and B would assume that the risk of disease being jointly exposed to A and B is different to the sum (addition) of the risk of being exposed to A only or exposed to B only. This model uses risk differences or differences between means. When interaction is found under this model, the findings indicate that the combined effect of A and B interact greater than additively (Table **2**). Similarly, under a multiplicative model the interaction between A and B would mean that the risk of disease if both A and B are present differs from the product of the independent effects of A and B (Table **2**).

Table 2. Statistical models used to study interactions between two risk factors, A and B (from Zammit *et al.* 2010 [29]).

Statistical Model	Relationship	Definition	Interaction
Additive	Additive	Risk (A and B) = Risk (A only) + Risk (B only) - Risk (Neither A nor B)	No (null hypothesis)
	Greater than additive	Risk (A and B) > Risk (A only) + Risk (B only) - Risk (Neither A nor B)	Yes
Multiplicative	Multiplicative	Risk ratio (A and B) = Risk ratio (A only) x Risk ratio (B only)	No (null hypothesis)
	Greater than multiplicative	Risk ratio (A and B) > Risk ratio (A only) x Risk ratio (B only)	Yes

Also, there is a particular type of interactions, called *qualitative* or *crossover* interactions, which are believed to be of major importance [29, 30]. In a qualitative interaction the effects go in opposite directions, for example, exposure to a particular environment is deleterious in carriers of a particular allele and protective in non-carriers and *vice-versa*. As opposite, in a quantitative or non-crossover interaction, there is a change in magnitude but not in direction of effects [30].

It is important to note that finding evidence for a gene-environment interaction effect does not provide *per se* information about the underlying biological or pathophysiological mechanism involved in the studied phenotype but can provide interesting data to progress in specific research lines.

GxE research has been a hot topic in fields related to human genetics in recent years, perhaps particularly so in psychiatry [19] but also in other complex diseases. However, in the last years, the enthusiasm for GxE research has been tempered by increasing scepticism [19, 22, 29, 31]. Dismissal about GxE studies has recently arisen mainly due to the failure to replicate, a problem that has also happened before with genetic association studies [31]. The lack of replication could be related to the high number of statistical tests that are possible when interaction effects are included in any analysis, increasing the risk of false positives [31]. Critics also worry about publication bias meaning the tendency to publish significant results more readily than non-significant ones [19]. Thus, the debate is not whether GxEs may be expected in disease but which GxE findings are meaningful and replicable and which are spurious and lead to wasted resources, false hope and increased scepticism.

6.3.1.1. Example 1: Gene-Environment Interaction in Mental Disorders: The Role of Childhood Maltreatment and MAOA Gene

From an evolutionary perspective, we can expect that gene-environment interaction findings would not only exist but may be quite common in behaviour and psychiatry genetics. Human development is a process in which individuals need to adapt to environmental changes. However, it is hard to believe that genetic variants are independent of the variation in individual responses to these environmental events, since these responses should be associated with individual differences in temperament, personality and psychophysiology, all of which are known to be under a certain degree of genetic influence [20].

The first evidence that genotype can regulate the capacity of a risk environment to induce mental disorders was reported in 2002 [21, 32]. These initial findings provided evidence that a functional polymorphism in the *MAOA* gene moderates the impact of early childhood maltreatment on the development of antisocial (AS) behaviour in males [32].

AS behaviour can be defined as a general pattern of disregard and violation of other individual's rights, which can begin at childhood or early adolescence and can continue into adulthood [33]. This behaviour is at the centre of violent delinquency nowadays, and is a matter of extreme social concern, to the point which the World Health Organisation has considered violence and its manifestations as one of the most important public health problems in the world [34].

From a genetic point of view, AS behaviour has been shown to demonstrate familial aggregation [35], with strong evidence for the heritability of the disorder coming from twin and adoption studies [36].

Because of the MAOA gene's ability to catabolise neurotransmitters, it is frequently a candidate gene in the study of behavioural traits and psychiatric diseases. In humans, a null allele at the MAOA locus was found in a Dutch family [37], with the affected males presenting no MAOA activity, since the MAOA gene is located in the X chromosome. In this family, the absence of MAOA (due to a point mutation in the 8th exon of the MAOA structural gene) conferred borderline mental retardation and abnormal behaviour (including impulsive aggression, arson, attempted rape, and exhibitionism) to all 5 affected males.

With respect to environmental risk factors, it has been classically shown that individuals that have suffered childhood maltreatment often present various forms of maladjustment, such as conduct disorder [38], substance abuse [39], aggressiveness [40, 41], AS behaviour [42, 43], delinquency, criminality, and violent behaviour [43 - 46], low levels of empathy [47], suicidal ideation and suicide attempts [44, 48].

In 2002, Caspi and collaborators showed the interacting effect of MAOA gene and childhood maltreatment in the appearance of antisocial behavior in males. Results showed that the development of conduct disorder, antisocial personality and adult violent crime occurred more frequently in children who suffered maltreatment and carried the genotype associated to low levels of MAOA expression than in those with a high-activity MAOA genotype [32] (Fig. **4**).

This publication contributed to change the view that gene-environment interactions (GxE) were rare and of limited importance, which had persisted from the 1980s to early 1990s in behavioural and psychiatric genetics [20]. In this period, it was assumed that genes would have relatively direct effects on disorder and that complex mental disorders would turn out to be caused by multiple different single gene conditions [49]. However, as Kendler pointed out, all the susceptibility genes for multifactorial disorders that have been discovered so far have been found to have very slight effects [50]. That initial approach also ignored the evidence from the rest of medicine that many risk factors operated on the basis of dimensional characteristics [51].

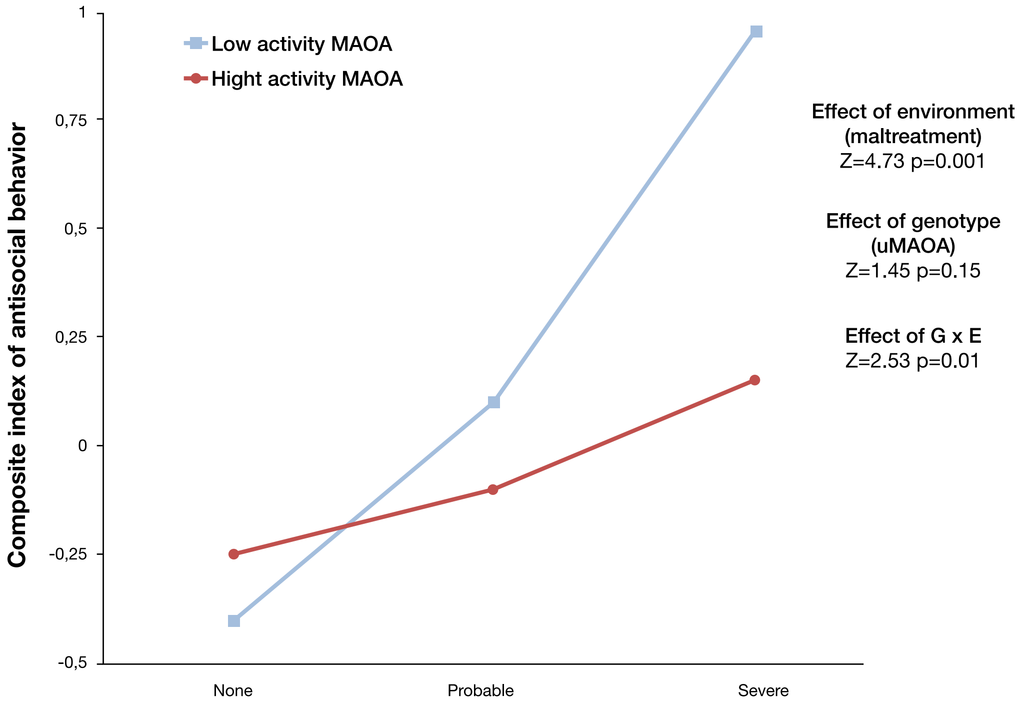


Fig. (4). Results of the interaction between *MAOA* activity and maltreatment on antisocial behavior, from the study by Caspi *et al.* 2002 [32]. For each individual (n=442), a composite antisocial index score was calculated (using the four measures of antisocial behavior evaluated in the study). The main effect of *MAOA* activity on the composite index of antisocial behavior was not significant, while the main effect of maltreatment was significant (p=0.001). A test of the interaction between *MAOA* activity and maltreatment showed a significant *G* x *E* interaction (p=0.01). This interaction within each genotype group revealed that the effect of childhood maltreatment on antisocial behavior was significantly weaker among males with high *MAOA* activity (p=0.03) than among males with low *MAOA* activity (p=0.001).

6.3.1.2. Example 2: Gene-Environment Interaction in Obesity

Obesity is one of the world's great health problems that contributes to chronic disease and premature mortality. Moreover, it is an increasing issue in childhood, and once developed, obesity is difficult to reverse (Fig. **5**).

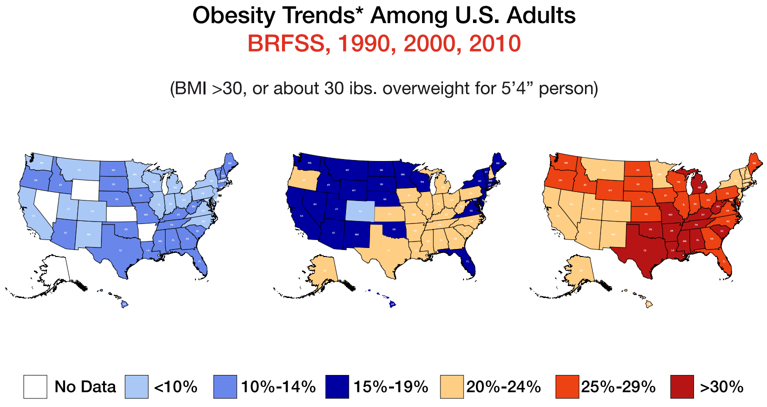

Obesity Trends* Among U.S. Adults
BRFSS, 1990, 2000, 2010
(BMI >30, or about 30 ibs. overweight for 5'4" person)
No Data
<10%
10%-14%
15%-19%
20%-24%
25%-29%
>30%

Fig. (5). Increase of Body Mass Index (BMI) in USA since 1990 to 2010.

To better understand its etiology, several evolutionary perspectives have been proposed [52 - 54]. The first hypothesis is known as the "thrifty gene" and proposes that the genetic predisposition to obesity was adaptive in the past, when having the possibility of storing large amounts of fat could have been selectively advantageous for survival [55]. The change of lifestyle from Paleolithic hunters-gatherers to agriculture, characterized by more sedentary occupations could be behind the prevalence of obesity and diabetes observed in modern societies. In the Paleolithic, genes promoting efficient fat accumulations would represent an advantage for humans in famine periods allowing survival [56]. On the contrary, nowadays, in those societies where food supply is always available, such genes are disadvantageous and the result is the widespread obesity [54, 56]. Neel's "thrifty gene hypothesis" has been useful to explain the high rates of obesity seen in certain populations after modernization, particularly Native Americans and many Pacific Islanders [57, 58].

A second hypothesis proposes that most mutations in the obesity susceptibility genes are neutral and have been drifting over evolutionary time [54, 59]. This new "drifty genes" hypothesis is a non-adaptive scenario and tries to provide an explanation for understanding why some individuals get obese while others remain resistant to obesity [54, 59]. Finally, Sellayah *et al.* (2014) [53] proposed the thermogenesis hypothesis, in which climatic selection has exerted a strong influence on genes. This work suggested that obesity could result from individual differences in brown adipose tissue (BAT) [53] and genes involved in the BAT thermogenic function could be positively selected to a better cold adaptation during the out of Africa migration. In warm climates individuals have a large surface area relative to body mass (*e.g.* slim, long trunk) that facilitate heat loss, whereas in cold climates individuals have a small surface area relative to body mass (*e.g.* bulky, short trunk) allowing heat retention. Epidemiological studies found different rates of obesity across certain ethnic groups; at the United States blacks, Hispanics, and people of Native American ancestry are more prone to develop an obese phenotype than European Caucasians and people of East Asian ancestry (Chinese, Japanese, and Koreans).

All these theories support the existence of a genetic basis to explain obesity. Quantitative genetic studies using family members (*e.g.* twins, adopted offspring) indicate that there is a high heritability for various body composition traits [60]. Several alleles have been identified as being associated with obesity, and the list of these alleles is increasing at a rapid pace. Two genes in particular have been identified as being associated with obesity risk: FTO (fat-mass and obesity-associated gene) and MC4R (melanocortin-4 receptor gene) [61, 62]. Nowadays, genome wide association studies (GWAS) have identified 97 genetic variants that are robustly associated with BMI [63]. The full 97 SNPs (Single Nucleotide Polymorphisms) can be combined into a polygenic obesity risk score that shows a quantitative association with body weight. However, to date, the known genetic variants collectively explain only a small fraction (2.7% of variation in BMI), a common problem also seen in other complex traits and diseases.

However, the rapid appearance of the obesity epidemic shows that it is not only caused by genetic changes; it is the environment that has changed rapidly. Then, these environmental changes accounting for the rapid increases in weight over the last 40, have conducted to the generation of the “obesogenic” environment term, which is intimately linked to changes in lifestyle and the food supply [64]. Developments in food production, processing, storage and preparation have resulted in highly-palatable and energy-dense foods becoming more accessible and cheaper [65]. Then, weight gain is related to an altered energy balance, with greater energy-intake than energy-expenditure. The prevalence of obesity today is just over double what it was in 1980 (increasing from 15% to 36%) but rates of morbid obesity (BMI>40) have increased more than four times (from 1.4% to 6.3%) [66, 67]. Moreover, it has been shown that weight gain has been greater at the upper end of the distribution in comparison to the central distribution [68, 69]. One theory for explaining this phenomenon is that the expression of “obesity risk genes” depends on the exposure to specific environments, which implies gene-environment interaction. This means that when environmental conditions are related to famine, individuals do no develop obesity, regardless the presence of genetic susceptibility; however, when conditions are related to abundance only those genetically susceptible individuals become obese (Fig. **6**).

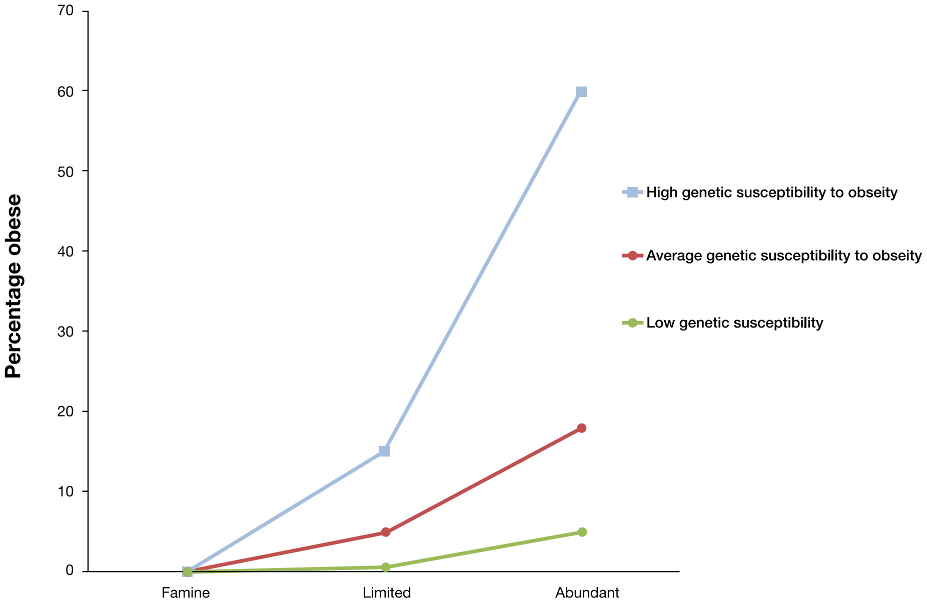


Fig. (6). A hypothetical GxE interaction model between the level of food supply and genetic susceptibility on the risk for obesity. In case of food abundance, those individuals with high genetic susceptibility will become obese (adapted from Llewellyn and Wardle 2015).

6.3.2. Gene-Environment Correlation

GxE has to be distinguished from another mechanism of gene-environment interplay, gene-environment correlation (Fig. **7**). Gene-environment correlation (rGE) refers to a situation where heritable factors are associated with specific environments. rGE can arise through environmental influences on genes, for example, through evolutionary processes such as natural selection. Of specific theoretical and practical interest in human complex traits, however, is the type of rGE that arises from genetic influences on the environment, that is, when an individual's genetically influenced traits (*e.g.*, behaviors, cognitive abilities, physical characteristics) in turn influence the kind of environment he or she will experience. In other words, detection of rGE effects indicates that the occurrence of the environmental exposure depends, at least partially, on genetic factors. A genetic variant may be associated with an increase of the likelihood that a person would be exposed to an environmental risk factor, which in turn, increases the likelihood to develop a particular trait or disease [24].

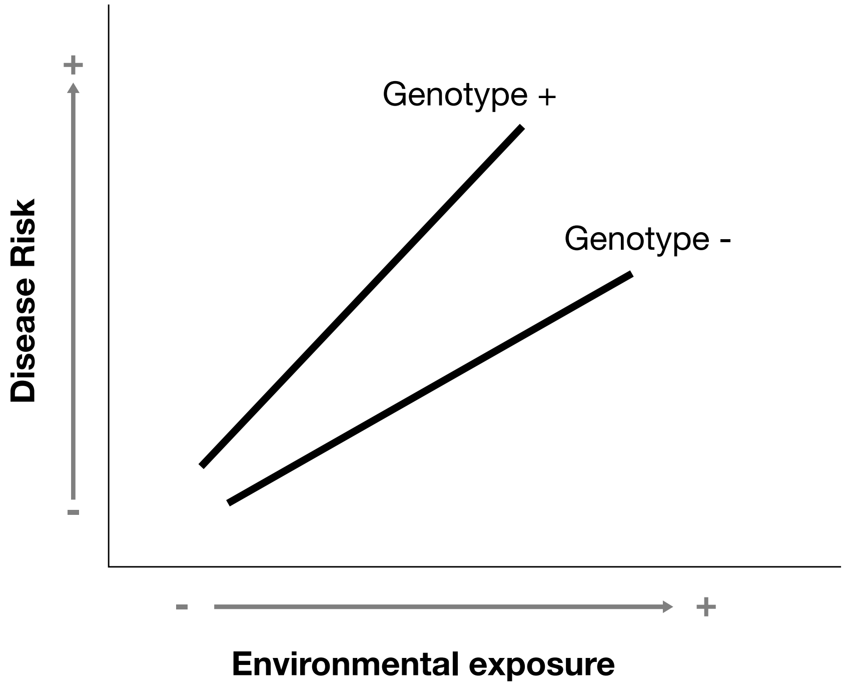


Fig. (7). Gene-environment correlation: the genetic control of exposure to the environment. The greater the genetic susceptibility, the higher the probability to be exposed to an environmental risk factor. Adapted from van Os and Marcelis 1998 [26].

Three types of rGE can be differentiated: passive, active and evocative rGE (Table **3**) [3]. Passive rGE refers to the association between the genotype a child inherits from her/his parents and the environment in which is raised. Passive rGE requires interactions between genetically related individuals. Evocative (or reactive) rGE refers to the association between an individual's genetically influenced behaviour and the reaction to that behaviour of those in the individual's environment. Active (or selective) rGE refers to the association between individual's genetically influenced traits or behaviours and the environmental niches selected by the individual.

Table 3. Three types of gene-environment correlation (from Plomin 2008 [8]).

Type	Description	Source of Environmental Influence
PASSIVE	Children receive genotypes correlated with their family environment	Parents and siblings
EVOCATIVE	Individuals are reacted to on the basis of their genetic propensities	Anybody
ACTIVE	Individuals seek or create environments correlated with their genetic proclivities	Anybody

For example, consider musical ability. If musical ability is heritable, musically talented children are likely to have parents with musical abilities who provide them with both genes and an environment leading to the development of musical ability (passive rGE). Musically gifted children might also be picked out at school and given special opportunities (evocative rGE). Even if one does anything about their musical talent, these children might seek out their own musical environments by selecting musical friends or otherwise creating musical experiences (active rGE) (Table **3**) [8].

There is evidence describing these types of rGE but these effects are still difficult to detect [20].

6.3.2.1. Example 1: The Role of Both GxE Interaction and Correlation in the Relationship between Cannabis and Schizophrenia

An interesting example, which in turn reflects the difficulty of differentiating between GxE mechanisms and rGE, is the case of cannabis and schizophrenia. On the one hand, cannabis use is a well-known risk factor for schizophrenia [70]. As expected in multifactorial complex diseases, cannabis use is not sufficient or necessary to cause disease; meaning that only a small proportion of cannabis users develop psychosis or that not all persons with a psychotic disorder have been exposed to cannabis. This suggests that the effects of using cannabis on the risk for psychosis vary across individuals depending on the distribution of other factors. Among these factors, the genetic variability observed in human populations is proposed to mediate the sensitivity to the cannabis effects [71, 72]. In this sense, genetically sensitive studies indicate that the risk of presenting cannabis-related psychotic experiences is much higher in at high risk individuals (*i.e.* individuals with psychosis liability measured psychometrically or defined as having an affected first degree relative) [73 - 75]. These results have leaded the view that the differential effect of cannabis may be related to individual predisposition to psychosis, which is strongly accepted to be influenced by genetic factors [76, 77]. Apart from these quantitative genetic studies, molecular studies have also reported evidence on gene-environment interaction regarding cannabis and schizophrenia. To this respect, several studies have explored the role of the single nucleotide polymorphism Val158Met (rs4680) in *COMT* gene on the differential risk to develop psychoses observed between cannabis users and non-users [78].

To briefly understand the interest of this gene, some aspects have to be mentioned. First, *COMT* gene encodes for Cathecol-O-Methyl Transferase, one of the several enzymes that degrade cathecolamines such as dopamine, epinephrine, and norepinephrine. Second, the Val158Met polymorphism alleles are associated with changes in COMT activity [79]. Third, the continued stimulation of cannabinoid receptors by tetrahydrocannabinol (THC, the psychoactive molecule of the plant Cannabis sativa) seems to increase the synthesis of dopamine in certain limbic structures or to decrease it in prefrontal structures and therefore modulate dopamine release [80, 81]. Then, studies have been based on the hypothesis that weak or subtle changes in COMT activity, may explain individual risk differences in the development of dopaminergic related psychotic symptoms in cannabis consumers.

The first study that provided evidence of interaction between cannabis and the Val158Met polymorphism was conducted by Caspi *et al.* [82]. This showed that adolescent cannabis use, but not onset of cannabis use in adulthood, was associated with a 10-fold increase in risk of developing psychotic disorder at age

26 among individuals homozygous for the Val allele, while cannabis use had no such effect on individuals homozygotes for the Met allele (Fig. **8**). These results, together with studies that have described the cannabis effects on the brain limited to adolescence both in human research and in experimental animal models [83, 84], have shed light on: i) the relationship between cannabis use, neurodevelopment processes and the emergence of psychiatric disorders, and ii) the sensitivity to cannabis effects may depend on the state of brain development and maturity at the moment of first exposure.

Not surprisingly, these results triggered several other studies with diverse designs aiming to replicate and to further understand the interaction between cannabis and *COMT* gene in relation to the risk for psychosis. However some studies have partially replicated the findings while others have not [85 - 90].

On the other hand, considering a gene-environment correlation scenario, family studies have shown that there is familial aggregation of cannabis use and twin studies have established that genetic factors are involved in the risk for cannabis abuse or dependence [91 - 94]. Moreover, it was reported that the risk for developing schizophrenia in first-degree relatives of patients with schizophrenia who used cannabis was ten times that of the families of patients who did not use cannabis [95]. At the molecular level, as an example, an association study based on the CNR1 gene (cannabinoid receptor type 1) described that patients with schizophrenia who also were cannabis users were carriers more frequently than expected by chance of a certain genotype than non-users [96]; suggesting the implication of this gene in cannabis exposure. Moreover, a recent Genome Wide Association Study has tested whether the cumulative burden of schizophrenia risk alleles carried by an individual predicts cannabis use [97]. Although directly predicting only a small amount of the variance, results from this study have led to the conclusion that psychosis liability by itself may explain part of the association between cannabis and psychosis, suggesting that there probably exists a bidirectional relationship between the risk factor and the disorder. Then, evidence seems to indicate that cannabis use and psychosis share at least part of their genetic roots and that both gene-environment interaction and correlation mechanisms should be better explored.

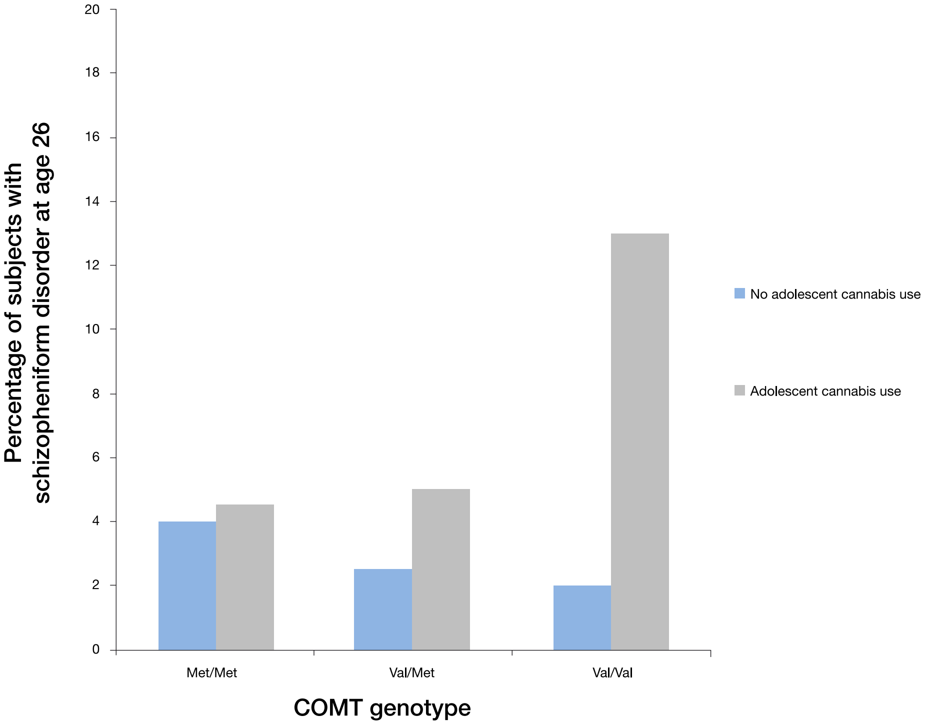


Fig. (8). Percentage of individuals meeting diagnostic criteria for schizophreniform disorder at age 26, as a function of COMT genotype (Val158Met) and adolescent-onset cannabis use, from the study by Caspi *et al.* 2005 [82]. The study included: n=199 Met/Met (151 cannabis non-users and 48 users), n=402 Val/Met (311 non-users and 91 users), and n=202 Val/Val (145 non-users and 54 users). In a hierarchical logistic regression model, the main effect of genotype was not significant, the main effect of adolescent cannabis exposure was significant (p=0.003), and the interaction between genotype and adolescent cannabis exposure was significant (p=0.025). Adolescent cannabis use was associated with increased risk of schizophreniform disorder in adulthood among Val/Val individuals (OR=10.9, 95% CI: 2.2–54.1) and, to a lesser extent, among Val/Met individuals (OR=2.5, 95% CI: 0.78 – 8.2), but not among Met/Met individuals.

6.3.2.2. Example 2: Gene-Environment Correlation: Quantitative and Molecular Genetics Data on the Role of Genetic Factors in Friendship Selection

Quantitative genetic studies are usually designed to estimate heritable and environmental influences on behavioural or other traits; however, this type of analysis can be also applied to aspects of the environment, such as peer relations. In humans, one of the most replicated findings in the social sciences is that people tend to associate with other people that they resemble [98, 99]. Although phenotypic similarity between friends might reflect only partially the effects of social influence [100, 101], genotypes are not materially susceptible to change. Therefore, evidence of genetic influences on friendships (or genotypic similarity between friends) may indicate an active rGE (*i.e.* may reflect an active selection of friends).

In this regard, recent studies including behavioural data of friends suggest that genetic factors can explain partly the process of friendship establishment. Data from the National Longitudinal Study of Adolescent Health (ADDHealth) was essential to arrive to this conclusion. In ADDHealth, a large subsample of adolescent sib-pairs (twin and non-twin) reported information on smoking and drinking for up to five male and five female school friends. Significant rGE was described: 64% of interindividual differences in affiliation with substance using friends could be attributed to genetic factors [102]. Similar rGE evidence was found with respect to friends' academic achievement, verbal intelligence, aggression, and depression in the ADDHealth sample [103].

It is interesting to note that the genetic effects on friendship relations are detected to be specific in adolescence (not observed in childhood) [104, 105], pointing towards an increased propensity for an active selection of friends based on heritable characteristics over the course of development [106].

Only few studies based on molecular genetics approaches have attempted to explain rGE mechanisms with respect to the friendship. Data from two independent samples, the ADDHealth study sample and the Framingham Heart Study Social Network (FHS) study, have been used to provide evidence for rGE in regard to peer selection [107]. Similarly to the ADDHealth, in the FHS study (a large population-based multi-generational cohort study) adult participants were asked to identify up to two close friends. All subjects (participants and the nominated friends) were genotyped for six genetic markers. Maps of the friendship networks showed clustering of genotypes and, after the application of strict controls for population stratification, the results indicated that, in both samples, friends were significantly similar with respect to the DRD2 (Dopamine Receptor D2 gene) genotype (*i.e.* friends presented the same genotype more frequently than expected by chance). It is remarkable that the minor allele of the studied genetic polymorphism at DRD2 has been associated with social alienation, antisocial behavior, and alcoholism [108 - 110]. Therefore, these genetic findings are in accordance with the previously mentioned quantitative genetic findings from the ADD Health study. The fact that friends tend to show certain degree of genetic similarity supports the idea that individuals may actively seek out friends with similar traits.

Globally, the description of gene–environment correlation phenomena both from quantitative and molecular genetics approaches has important implications for social and genetic epidemiology. These processes may also have significance for the study of social factors related to health and health-related behaviours.

CONCLUDING REMARKS

The dichotomy nature-nurture or the innate *versus* the acquired (that prevailed in the genetics of the XXth century) has given way to the acceptance of the multifactorial origin of most of the human traits. Then, nowadays it is widely accepted that the individual genetic background may constitute an element in continuous interaction with the environment.

A huge number of quantitative genetics studies have firmly established the involvement of genetic and environmental factors in the etiology of human traits, behaviors and diseases. Molecular genetics studies and epidemiological studies have joined efforts to describe specific genes and/or environments involved in the origin of complex traits. Traditionally, GxE interactions were investigated using candidate-gene or candidate-genetic pathways studies.

Methodological advances in molecular genetics have allowed designing new molecular approaches based on the whole genome such as next-generation sequencing or genome wide analyses based in common variants (GWAS-Genome Wide Association Studies). In the current post-GWAS era, the focus is on: i) integrating findings from the vast body of data that has been generated through large consortia, and ii) collecting environmental data from those epidemiological studies with well-characterized exposure information included in the consortia. A key feature of the next phase should be a renewed focus on G×E (GEWIS- Gene Environment-Wide Interaction Studies), but this will require careful consideration of epidemiologic study design, exposure assessment and methods of analysis, with particular attention to harmonization of these features across the consortia.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Supported by: This study was supported by: i) The Network of European Funding for Neuroscience Research, ERA-NET NEURON (PiM2010ERN-00642), iii) Instituto de Salud Carlos III through the projects PI12/00018 and PI15/01420 (co-funded by European Regional Development Fund /European Social Fund, "Investing in your future"). Thanks to the Comissionat per a Universitats i Recerca del DIUE (*2014SGR1636*).

REFERENCES

[1] Turkheimer E, Waldron M. Nonshared environment: a theoretical, methodological, and quantitative review. Psychol Bull 2000; 126(1): 78-108.
[http://dx.doi.org/10.1037/0033-2909.126.1.78] [PMID: 10668351]

[2] Wachs TD. The use and abuse of environment in behavior-genetic research. Child Dev 1983; 54(2): 396-407.
[http://dx.doi.org/10.2307/1129700] [PMID: 6872631]

[3] Plomin R, DeFries JC, Loehlin JC. Genotype-environment interaction and correlation in the analysis of human behavior. Psychol Bull 1977; 84(2): 309-22.
[http://dx.doi.org/10.1037/0033-2909.84.2.309] [PMID: 557211]

[4] Scarr S, McCartney K. How people make their own environments: a theory of genotype greater than environment effects. Child Dev 1983; 54(2): 424-35.
[PMID: 6683622]

[5] Fisher RA. The correlation between relatives under the supposition of Mendelian inheritance. Trans Roy Soc 1918; 52: 399-433.
[http://dx.doi.org/10.1017/S0080456800012163]

[6] Brüne M. Textbook of evolutionary psychiatry: The origins of psychopathology. Oxford: Oxford University Press 2010.

[7] Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nat Rev Genet 2002; 3(11): 872-82.
[http://dx.doi.org/10.1038/nrg932] [PMID: 12415317]

[8] Plomin R, DeFries JC, McClearn G, McGuffin P. Behavioral Genetics. 5th ed., New York: Worth Publishers 2008.

[9] Hall JG. Twinning. Lancet 2003; 362(9385): 735-43.
[http://dx.doi.org/10.1016/S0140-6736(03)14237-7] [PMID: 12957099]

[10] Fuster V, Zuluaga P, Colantonio S, de Blas C. Factors associated with recent increase of multiple births in Spain. Twin Res Hum Genet 2008; 11(1): 70-6.
[http://dx.doi.org/10.1375/twin.11.1.70] [PMID: 18251678]

[11] Pison G, DAddato AV. Frequency of twin births in developed countries. Twin Res Hum Genet 2006; 9(2): 250-9.
[http://dx.doi.org/10.1375/twin.9.2.250] [PMID: 16611495]

[12] van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Rev Genet 2012; 13(9): 640-53.
[http://dx.doi.org/10.1038/nrg3243] [PMID: 22847273]

[13] Rijsdijk FV, Sham PC. Analytic approaches to twin data using structural equation models. Brief Bioinform 2002; 3(2): 119-33.
[http://dx.doi.org/10.1093/bib/3.2.119] [PMID: 12139432]

[14] Plomin R, Owen MJ, McGuffin P. The genetic basis of complex human behaviors. Science 1994; 264(5166): 1733-9.
[http://dx.doi.org/10.1126/science.8209254] [PMID: 8209254]

[15] Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. Parental treatment and the equal environment assumption in twin studies of psychiatric illness. Psychol Med 1994; 24(3): 579-90.
[http://dx.doi.org/10.1017/S0033291700027732] [PMID: 7991740]

[16] Derks EM, Dolan CV, Boomsma DI. A test of the equal environment assumption (EEA) in multivariate twin studies. Twin Res Hum Genet 2006; 9(3): 403-11.
[http://dx.doi.org/10.1375/twin.9.3.403] [PMID: 16790150]

[17] Evans DM, Martin NG. The validity of twin studies. GeneScreen 2000; 1: 77-9.
[http://dx.doi.org/10.1046/j.1466-9218.2000.00027.x]

[18] Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. Nat Rev Genet 2009; 10(12): 872-8.
[http://dx.doi.org/10.1038/nrg2670] [PMID: 19859063]

[19] Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry 2011; 168(10): 1041-9.
[http://dx.doi.org/10.1176/appi.ajp.2011.11020191] [PMID: 21890791]

[20] Rutter M, Moffitt TE, Caspi A. Gene-environment interplay and psychopathology: multiple varieties but real effects. J Child Psychol Psychiatry 2006; 47(3-4): 226-61.
[http://dx.doi.org/10.1111/j.1469-7610.2005.01557.x] [PMID: 16492258]

[21] Caspi A, Moffitt TE. Gene-environment interactions in psychiatry: joining forces with neuroscience. Nat Rev Neurosci 2006; 7(7): 583-90.
[http://dx.doi.org/10.1038/nrn1925] [PMID: 16791147]

[22] Eaves LJ. The resolution of genotype x environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1984; 1(3): 215-28.
[http://dx.doi.org/10.1002/gepi.1370010302] [PMID: 6544238]

[23] Rutter M, Silberg J. Gene-environment interplay in relation to emotional and behavioral disturbance. Annu Rev Psychol 2002; 53: 463-90.
[http://dx.doi.org/10.1146/annurev.psych.53.100901.135223] [PMID: 11752493]

[24] van Os J, Sham P. Gene-environment correlation and interaction in schizophrenia. In: Murray RM, Jones PB, Susser E, van Os J, Cannon M, Eds. The Epidemiology of Schizophrenia. Cambridge: Cambridge University Press 2003.

[25] Belsky J, Jonassaint C, Pluess M, Stanton M, Brummett B, Williams R. Vulnerability genes or plasticity genes? Mol Psychiatry 2009; 14(8): 746-54.
[http://dx.doi.org/10.1038/mp.2009.44] [PMID: 19455150]

[26] van Os J, Marcelis M. The ecogenetics of schizophrenia: a review. Schizophr Res 1998; 32(2): 127-35.
[http://dx.doi.org/10.1016/S0920-9964(98)00049-8] [PMID: 9713909]

[27] Thomas D. Geneenvironment-wide association studies: emerging approaches. Nat Rev Genet 2010; 11(4): 259-72.
[http://dx.doi.org/10.1038/nrg2764] [PMID: 20212493]

[28] van Winkel R, Stefanis NC, Myin-Germeys I. Psychosocial stress and psychosis. A review of the neurobiological mechanisms and the evidence for gene-stress interaction. Schizophr Bull 2008; 34(6): 1095-105.
[http://dx.doi.org/10.1093/schbul/sbn101] [PMID: 18718885]

[29] Zammit S, Owen MJ, Lewis G. Misconceptions about gene-environment interactions in psychiatry. Evid Based Ment Health 2010; 13(3): 65-8.
[http://dx.doi.org/10.1136/ebmh1056] [PMID: 20682811]

[30] Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 1985; 41(2): 361-72.
[http://dx.doi.org/10.2307/2530862] [PMID: 4027319]

[31] Munafò MR, Flint J. Replication and heterogeneity in gene x environment interaction studies. Int J Neuropsychopharmacol 2009; 12(6): 727-9.
[http://dx.doi.org/10.1017/S1461145709000479] [PMID: 19476681]

[32] Caspi A, McClay J, Moffitt TE, *et al.* Role of genotype in the cycle of violence in maltreated children. Science 2002; 297(5582): 851-4.
[http://dx.doi.org/10.1126/science.1072290] [PMID: 12161658]

[33] APA Diagnostic and Statistical Manual of Mental Disorders. Fourth Edition, Text Revision., Washington, DC: American Psychiatric Association 2000.

[34] World report on violence and health. Geneva: World Health Organisation 2002.

[35] Cloninger CR, Christiansen KO, Reich T, Gottesman II. Implications of sex differences in the prevalences of antisocial personality, alcoholism, and criminality for familial transmission. Arch Gen Psychiatry 1978; 35(8): 941-51.
[http://dx.doi.org/10.1001/archpsyc.1978.01770320035002] [PMID: 354554]

[36] McGuffin PG. Genetic influences on normal and abnormal development. In: Rutter MH, Ed. Child psychiatry: modern approaches. London: Blackwell 1984.

[37] Brunner HG, Nelen M, Breakefield XO, Ropers HH, van Oost BA. Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. Science 1993; 262(5133): 578-80. [http://dx.doi.org/10.1126/science.8211186] [PMID: 8211186]

[38] Friedrich WN, Luecke WJ. Young school-age sexually aggressive-children. Prof Psychol Res Pr 1988; 19: 155-64. [http://dx.doi.org/10.1037/0735-7028.19.2.155]

[39] Herman JL. Histories of violence in an outpatient population: an exploratory study. Am J Orthopsychiatry 1986; 56(1): 137-41. [http://dx.doi.org/10.1111/j.1939-0025.1986.tb01550.x] [PMID: 3946562]

[40] Graybill JR, Mackie DJ, House AE. Aggression in college-students who were abused as children. J Coll Student Dev 1985; 26: 492-5.

[41] Zeiller B. [Physical and psychological abuse: follow-up of abused and delinquent adolescents]. Child Abuse Negl 1982; 6(2): 207-10. [http://dx.doi.org/10.1016/0145-2134(82)90015-1] [PMID: 6892303]

[42] Christozov C, Toteva S. Abuse and neglect of children brought up in families with an alcoholic father in Bulgaria. Child Abuse Negl 1989; 13(1): 153-5. [http://dx.doi.org/10.1016/0145-2134(89)90039-2] [PMID: 2706558]

[43] Oliver JE. Successive generations of child maltreatment. The children. Br J Psychiatry 1988; 153: 543-53. [http://dx.doi.org/10.1192/bjp.153.4.543] [PMID: 3250697]

[44] Cavaiola AA, Schiff M. Behavioral sequelae of physical and/or sexual abuse in adolescents. Child Abuse Negl 1988; 12(2): 181-8. [http://dx.doi.org/10.1016/0145-2134(88)90026-9] [PMID: 3395894]

[45] Widom CS. Child abuse, neglect, and adult behavior: research design and findings on criminality, violence, and child abuse. Am J Orthopsychiatry 1989; 59(3): 355-67. [http://dx.doi.org/10.1111/j.1939-0025.1989.tb01671.x] [PMID: 2764070]

[46] Widom CS. The cycle of violence. Science 1989; 244(4901): 160-6. [http://dx.doi.org/10.1126/science.2704995] [PMID: 2704995]

[47] Miller PA, Eisenberg N. The relation of empathy to aggressive and externalizing/antisocial behavior. Psychol Bull 1988; 103(3): 324-44. [http://dx.doi.org/10.1037/0033-2909.103.3.324] [PMID: 3289071]

[48] Green AH. Self-destructive behavior in battered children. Am J Psychiatry 1978; 135(5): 579-82. [http://dx.doi.org/10.1176/ajp.135.5.579] [PMID: 645951]

[49] Kidd KK. Trials and tribulations in the search for genes causing neuropsychiatric disorders. Soc Biol 1991; 38(3-4): 163-78. [PMID: 1801198]

[50] Kendler KS. A gene for...: the nature of gene action in psychiatric disorders. Am J Psychiatry 2005;

162(7): 1243-52.
[http://dx.doi.org/10.1176/appi.ajp.162.7.1243] [PMID: 15994704]

[51] Rutter M. Categories, dimensions and the mental health of children and adolescents. In: King JA, Ferris CF, Lederhendler II, Eds. Roots of mental illness in children. New York: The New York Academy of Sciences 2003.
[http://dx.doi.org/10.1196/annals.1301.002]

[52] Albuquerque D, Stice E, Rodríguez-López R, Manco L, Nóbrega C. Current review of genetics of human obesity: from molecular mechanisms to an evolutionary perspective. Mol Genet Genomics 2015; 290(4): 1191-221.
[http://dx.doi.org/10.1007/s00438-015-1015-9] [PMID: 25749980]

[53] Sellayah D, Cagampang FR, Cox RD. On the evolutionary origins of obesity: a new hypothesis. Endocrinology 2014; 155(5): 1573-88.
[http://dx.doi.org/10.1210/en.2013-2103] [PMID: 24605831]

[54] Speakman JR. Evolutionary perspectives on the obesity epidemic: adaptive, maladaptive, and neutral viewpoints. Annu Rev Nutr 2013; 33: 289-317.
[http://dx.doi.org/10.1146/annurev-nutr-071811-150711] [PMID: 23862645]

[55] Neel JV. Diabetes mellitus: a thrifty genotype rendered detrimental by progress? Am J Hum Genet 1962; 14: 353-62.
[PMID: 13937884]

[56] Speakman JR. Thrifty genes for obesity and the metabolic syndrometime to call off the search? Diab Vasc Dis Res 2006; 3(1): 7-11.
[http://dx.doi.org/10.3132/dvdr.2006.010] [PMID: 16784175]

[57] Baker PT. Migration, genetics, and the degenerative diseases of South Pacific islanders. In: Boyce AJ, Ed. Migration and Mobility. London: Taylor and Francis 1984; pp. 209-39.

[58] Szathmary EJ. Non-insulin dependent diabetes mellitus among aboriginal North Americans. Annu Rev Anthropol 1994; 111: 263-81.

[59] Speakman JR. Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the drifty gene hypothesis. Int J Obes 2008; 32(11): 1611-7.
[http://dx.doi.org/10.1038/ijo.2008.161] [PMID: 18852699]

[60] Owen JB. Genetic aspects of body composition. Nutrition 1999; 15(7-8): 609-13.
[http://dx.doi.org/10.1016/S0899-9007(99)00097-0] [PMID: 10422098]

[61] Frayling TM, Timpson NJ, Weedon MN, *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 2007; 316(5826): 889-94.
[http://dx.doi.org/10.1126/science.1141634] [PMID: 17434869]

[62] Loos RJ, Lindgren CM, Li S, *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 2008; 40(6): 768-75.
[http://dx.doi.org/10.1038/ng.140] [PMID: 18454148]

[63] Locke AE, Kahali B, Berndt SI, *et al.* Genetic studies of body mass index yield new insights for obesity biology. Nature 2015; 518(7538): 197-206.
[http://dx.doi.org/10.1038/nature14177] [PMID: 25673413]

[64] Llewellyn C, Wardle J. Behavioral susceptibility to obesity: Gene-environment interplay in the development of weight. Physiol Behav 2015; 152(Pt B): 494-501.

[65] Swinburn BA, Sacks G, Hall KD, *et al.* The global obesity pandemic: shaped by global drivers and local environments. Lancet 2011; 378(9793): 804-14.
[http://dx.doi.org/10.1016/S0140-6736(11)60813-1] [PMID: 21872749]

[66] Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among US adults, 19992010. JAMA 2012; 307(5): 491-7.
[http://dx.doi.org/10.1001/jama.2012.39] [PMID: 22253363]

[67] Ogden CL, Carroll MD, Curtin LR, Lamb MM, Flegal KM. Prevalence of high body mass index in US children and adolescents, 20072008. JAMA 2010; 303(3): 242-9.
[http://dx.doi.org/10.1001/jama.2009.2012] [PMID: 20071470]

[68] Ogden CL, Yanovski SZ, Carroll MD, Flegal KM. The epidemiology of obesity. Gastroenterology 2007; 132(6): 2087-102.
[http://dx.doi.org/10.1053/j.gastro.2007.03.052] [PMID: 17498505]

[69] Wardle J, Boniface D. Changes in the distributions of body mass index and waist circumference in English adults, 1993/1994 to 2002/2003. Int J Obes 2008; 32(3): 527-32.
[http://dx.doi.org/10.1038/sj.ijo.0803740] [PMID: 17923859]

[70] Henquet C, Murray R, Linszen D, van Os J. The environment and schizophrenia: the role of cannabis use. Schizophr Bull 2005; 31(3): 608-12.
[http://dx.doi.org/10.1093/schbul/sbi027] [PMID: 15976013]

[71] Henquet C, Di Forti M, Morrison P, Kuepper R, Murray RM. Gene-environment interplay between cannabis and psychosis. Schizophr Bull 2008; 34(6): 1111-21.
[http://dx.doi.org/10.1093/schbul/sbn108] [PMID: 18723841]

[72] DSouza DC, Sewell RA, Ranganathan M. Cannabis and psychosis/schizophrenia: human studies. Eur Arch Psychiatry Clin Neurosci 2009; 259(7): 413-31.
[http://dx.doi.org/10.1007/s00406-009-0024-2] [PMID: 19609589]

[73] Verdoux H, Gindre C, Sorbara F, Tournier M, Swendsen JD. Effects of cannabis and psychosis vulnerability in daily life: an experience sampling test study. Psychol Med 2003; 33(1): 23-32.
[http://dx.doi.org/10.1017/S0033291702006384] [PMID: 12537033]

[74] Henquet C, Krabbendam L, Spauwen J, *et al.* Prospective cohort study of cannabis use, predisposition for psychosis, and psychotic symptoms in young people. BMJ 2005; 330(7481): 11.
[http://dx.doi.org/10.1136/bmj.38267.664086.63] [PMID: 15574485]

[75] van Os J, Bak M, Hanssen M, Bijl RV, de Graaf R, Verdoux H. Cannabis use and psychosis: a longitudinal population-based study. Am J Epidemiol 2002; 156(4): 319-27.
[http://dx.doi.org/10.1093/aje/kwf043] [PMID: 12181101]

[76] Lichtenstein P, Yip BH, Björk C, *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 2009; 373(9659): 234-9.
[http://dx.doi.org/10.1016/S0140-6736(09)60072-6] [PMID: 19150704]

[77] Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 2003; 60(12): 1187-92.

[http://dx.doi.org/10.1001/archpsyc.60.12.1187] [PMID: 14662550]

[78] Di Forti M, Sallis H, Allegri F, *et al.* Daily use, especially of high-potency cannabis, drives the earlier onset of psychosis in cannabis users. Schizophr Bull 2014; 40(6): 1509-17.
[http://dx.doi.org/10.1093/schbul/sbt181] [PMID: 24345517]

[79] Lotta T, Vidgren J, Tilgmann C, *et al.* Kinetics of human soluble and membrane-bound catechol O-methyltransferase: a revised mechanism and description of the thermolabile variant of the enzyme. Biochemistry 1995; 34(13): 4202-10.
[http://dx.doi.org/10.1021/bi00013a008] [PMID: 7703232]

[80] Cheer JF, Wassum KM, Heien ML, Phillips PE, Wightman RM. Cannabinoids enhance subsecond dopamine release in the nucleus accumbens of awake rats. J Neurosci 2004; 24(18): 4393-400.
[http://dx.doi.org/10.1523/JNEUROSCI.0529-04.2004] [PMID: 15128853]

[81] Pertwee RG. Inverse agonism and neutral antagonism at cannabinoid CB1 receptors. Life Sci 2005; 76(12): 1307-24.
[http://dx.doi.org/10.1016/j.lfs.2004.10.025] [PMID: 15670612]

[82] Caspi A, Moffitt TE, Cannon M, *et al.* Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. Biol Psychiatry 2005; 57(10): 1117-27.
[http://dx.doi.org/10.1016/j.biopsych.2005.01.026] [PMID: 15866551]

[83] Bossong MG, Niesink RJ. Adolescent brain maturation, the endogenous cannabinoid system and the neurobiology of cannabis-induced schizophrenia. Prog Neurobiol 2010; 92(3): 370-85.
[http://dx.doi.org/10.1016/j.pneurobio.2010.06.010] [PMID: 20624444]

[84] Schneider M, Koch M. Chronic pubertal, but not adult chronic cannabinoid treatment impairs sensorimotor gating, recognition memory, and the performance in a progressive ratio task in adult rats. Neuropsychopharmacology 2003; 28(10): 1760-9.
[http://dx.doi.org/10.1038/sj.npp.1300225] [PMID: 12888772]

[85] Estrada G, Fatjó-Vilas M, Muñoz MJ, *et al.* Cannabis use and age at onset of psychosis: further evidence of interaction with COMT Val158Met polymorphism. Acta Psychiatr Scand 2011; 123(6): 485-92.
[http://dx.doi.org/10.1111/j.1600-0447.2010.01665.x] [PMID: 21231925]

[86] Zammit S, Owen MJ, Evans J, Heron J, Lewis G. Cannabis, COMT and psychotic experiences. Br J Psychiatry 2011; 199(5): 380-5.
[http://dx.doi.org/10.1192/bjp.bp.111.091421] [PMID: 21947654]

[87] Zammit S, Spurlock G, Williams H, *et al.* Genotype effects of CHRNA7, CNR1 and COMT in schizophrenia: interactions with tobacco and cannabis use. Br J Psychiatry 2007; 191: 402-7.
[http://dx.doi.org/10.1192/bjp.bp.107.036129] [PMID: 17978319]

[88] Costas J, Sanjuán J, Ramos-Ríos R, *et al.* Interaction between COMT haplotypes and cannabis in schizophrenia: a case-only study in two samples from Spain. Schizophr Res 2011; 127(1-3): 22-7.
[http://dx.doi.org/10.1016/j.schres.2011.01.014] [PMID: 21310591]

[89] De Sousa KR, Tiwari AK, Giuffra DE, Mackenzie B, Zai CC, Kennedy JL. Age at onset of schizophrenia: cannabis, COMT gene, and their interactions. Schizophr Res 2013; 151(1-3): 289-90.

[http://dx.doi.org/10.1016/j.schres.2013.10.037] [PMID: 24268936]

[90] Pelayo-Terán JM, Pérez-Iglesias R, Mata I, Carrasco-Marín E, Vázquez-Barquero JL, Crespo-Facorro B. Catechol-O-Methyltransferase (COMT) Val158Met variations and cannabis use in first-episode non-affective psychosis: clinical-onset implications. Psychiatry Res 2010; 179(3): 291-6. [http://dx.doi.org/10.1016/j.psychres.2009.08.022] [PMID: 20493536]

[91] Tsuang MT, Lyons MJ, Eisen SA, *et al.* Genetic influences on DSM-III-R drug abuse and dependence: a study of 3,372 twin pairs. Am J Med Genet 1996; 67(5): 473-7. [http://dx.doi.org/10.1002/(SICI)1096-8628(19960920)67:5<473::AID-AJMG6>3.0.CO;2-L] [PMID: 8886164]

[92] Merikangas KR, Stolar M, Stevens DE, *et al.* Familial transmission of substance use disorders. Arch Gen Psychiatry 1998; 55(11): 973-9. [http://dx.doi.org/10.1001/archpsyc.55.11.973] [PMID: 9819065]

[93] Agrawal A, Lynskey MT. The genetic epidemiology of cannabis use, abuse and dependence. Addiction 2006; 101(6): 801-12. [http://dx.doi.org/10.1111/j.1360-0443.2006.01399.x] [PMID: 16696624]

[94] Kendler KS, Schmitt E, Aggen SH, Prescott CA. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. Arch Gen Psychiatry 2008; 65(6): 674-82. [http://dx.doi.org/10.1001/archpsyc.65.6.674] [PMID: 18519825]

[95] McGuire PK, Jones P, Harvey I, Williams M, McGuffin P, Murray RM. Morbid risk of schizophrenia for relatives of patients with cannabis-associated psychosis. Schizophr Res 1995; 15(3): 277-81. [http://dx.doi.org/10.1016/0920-9964(94)00053-B] [PMID: 7632625]

[96] Leroy S, Griffon N, Bourdel MC, Olié JP, Poirier MF, Krebs MO. Schizophrenia and the cannabinoid receptor type 1 (CB1): association study using a single-base polymorphism in coding exon 1. Am J Med Genet 2001; 105(8): 749-52. [http://dx.doi.org/10.1002/ajmg.10038] [PMID: 11803524]

[97] Power RA, Verweij KJ, Zuhair M, *et al.* Genetic predisposition to schizophrenia associated with increased use of cannabis. Mol Psychiatry 2014; 19(11): 1201-4. [http://dx.doi.org/10.1038/mp.2014.51] [PMID: 24957864]

[98] McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annu Rev Sociol 2001; 27: 415-44. [http://dx.doi.org/10.1146/annurev.soc.27.1.415]

[99] Zeng Z, Xie Y. A preference-opportunity-choice framework with applications to intergroup friendship. Am J Sociol 2008; 114(3): 615-48. [http://dx.doi.org/10.1086/592863] [PMID: 19569394]

[100] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med 2007; 357(4): 370-9. [http://dx.doi.org/10.1056/NEJMsa066082] [PMID: 17652652]

[101] Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. N Engl J Med 2008; 358(21): 2249-58.

[http://dx.doi.org/10.1056/NEJMsa0706154] [PMID: 18499567]

[102] Cleveland HH, Wiebe RP, Rowe DC. Sources of exposure to smoking and drinking friends among adolescents: a behavioral-genetic evaluation. J Genet Psychol 2005; 166(2): 153-69.
[PMID: 15906929]

[103] Guo JF, Kuang Yang Y, Tsing Chiu N, *et al.* The correlation between striatal dopamine D2/D3 receptor availability and verbal intelligence quotient in healthy volunteers. Psychol Med 2006; 36(4): 547-54.
[http://dx.doi.org/10.1017/S0033291705006732] [PMID: 16359604]

[104] Rose AJ. Co-rumination in the friendships of girls and boys. Child Dev 2002; 73(6): 1830-43.
[http://dx.doi.org/10.1111/1467-8624.00509] [PMID: 12487497]

[105] Bullock BM, Deater-Deckard K, Leve LD. Deviant peer affiliation and problem behavior: a test of genetic and environmental influences. J Abnorm Child Psychol 2006; 34(1): 29-41.
[http://dx.doi.org/10.1007/s10802-005-9004-9] [PMID: 16550453]

[106] Bredgen M. Genetics and peer relations: a review. J Res Adolesc 2012; 22(3): 410-37.

[107] Fowler JH, Settle JE, Christakis NA. Correlated genotypes in friendship networks. Proc Natl Acad Sci USA 2011; 108(5): 1993-7.
[http://dx.doi.org/10.1073/pnas.1011687108] [PMID: 21245293]

[108] Hill SY, Zezza N, Wipprecht G, Locke J, Neiswanger K. Personality traits and dopamine receptors (D2 and D4): linkage studies in families of alcoholics. Am J Med Genet 1999; 88(6): 634-41.
[http://dx.doi.org/10.1002/(SICI)1096-8628(19991215)88:6<634::AID-AJMG11>3.0.CO;2-M]
[PMID: 10581482]

[109] Le Foll B, Gallo A, Le Strat Y, Lu L, Gorwood P. Genetics of dopamine receptors and drug addiction: a comprehensive review. Behav Pharmacol 2009; 20(1): 1-17.
[http://dx.doi.org/10.1097/FBP.0b013e3283242f05] [PMID: 19179847]

[110] Ponce G, Jimenez-Arriero MA, Rubio G, *et al.* The A1 allele of the DRD2 gene (TaqI A polymorphisms) is associated with antisocial personality in a sample of alcohol-dependent patients. Eur Psychiatry 2003; 18(7): 356-60.
[http://dx.doi.org/10.1016/j.eurpsy.2003.06.006] [PMID: 14643564]

CHAPTER 7

Ancient DNA: From Single Words to Full Libraries in 30 Years

Marc Simón and **Assumpció Malgosa**[*]

Departament de Biologia Animal, Biologia Vegetal i Ecologia, Unitat d'Antropologia Biològica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Abstract: Since the arrival of the technology that permitted to recover and study ancient genetic material 30 years ago, its success has enjoyed steady growth, providing answers to a huge variety of fields, from personal identification to a better understanding of ancient human behavior, as well as the intricated evolution of our species or the recovery of genetic material from extinct ones. However, this field has also been accompanied by some handicaps which have complicated its improvement, as the damage that the individuals may have suffered over time and most notoriously contamination. A brief synthesis of the principal landmarks in this field's history and the steps taken to overcome these problems are exposed in detail.

Keywords: Ancient diseases, Ancient DNA, Contamination, Damage, Diagenesis, Endogenous DNA, Extinct species, Forensic identification, Human history, Identification, Kinship, PCR, Putrefaction, Storing conditions.

7.1. HISTORY OF ANCIENT DNA ANALYSIS

Since the first study that documented the retrieval of ancient DNA (aDNA) was published in 1980 [1] widely unknown as it was published in Chinese, the recovery of mitochondrial DNA (mtDNA) from the extinct Equus quagga in 1984 [2] and of nuclear DNA (nuDNA) from a 2.400-year-old Egyptian mummy [3], many advances have been made in paleogenetics. Owing to this, many works that just used physical data have been now complemented with genetic analyses.

[*] **Corresponding author Assumpció Malgosa:** Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain; Tel/Fax: +34935811860; E-mail: assumpcio.malgosa@uab.cat

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

The most important step that propelled these improvements was the advent of the polymerase chain reaction (PCR) discovered by Mullis in 1983 [4, 5]. PCR could amplify previously selected DNA fragments up to a level that permitted direct sequencing, starting from very tiny quantities of this molecule, sometimes as a single one, also diminishing the time needed to retrieve useful information.

The first application of the PCR to ancient genetic material was carried out by Pääbo and Wilson in 1988 [6]. In this and subsequent studies, it was seen that to analyze DNA from ancient samples, application of this technique was an almost unavoidable requisite due to the low quantity of DNA extracted [7, 8]. Moreover, PCR solved some of the problems caused by low cloning efficiencies [7]. Using it, different groups could amplify and sequence DNA from soft tissues of natural or artificial human mummies as well as from recently extinct animal species (for example *Thylacinus cynocephalus* [9]).

However, these tissues represented very rare and geographically restricted remains whose conservation was strictly dependent on particular conditions. In addition, the fact that bones or teeth are the remains most currently found in archeological contexts prompted scientists to study the presence of DNA in these samples. In this sense, three different studies came out in 1989 reporting the recovery of genetic material from bone [8, 10, 11] encompassing antiquities ranging from 60 to 3,500 BP. In 1990 Hanni and co-workers retrieved mtDNA fragments from teeth and bones of individuals ranging from 150 to 5.500 years BP [12]. They amplified a specific DNA fragment of 121 base pairs (bp) of human mtDNA, which was also cloned and sequenced in the most recent bone.

In 1991, a major milestone was achieved by Dr. Erika Hagelberg [13], who was responsible for the first forensic identification of a murder victim accepted by an English court using a skeletal genetic analysis. Two years later, teeth were used for the first time to check for the possible kinship of two sets of individuals from the 1st and 5th centuries AD using mtDNA and short tandem repeats (STRs) [14].

As paleogeneticists realized that it was imperative to know under which conditions such experiments were optimized, different assays to check for the best conditions to carry them out appeared during those years. For instance, in 1994

Woodward and co-workers [15] compared the ease of extraction, resulting quality of the DNA and ultimate reliability of data obtained from soft tissue and from teeth of ancient Egyptian mummies from a cemetery of the Greco-Roman period (200 BC-800 AD), showing that teeth were a better choice for recovering aDNA. Likewise, in 1999, Burger and co-workers [16] investigated teeth from 18 individuals from three archeological sites of similar age (1st and 2nd millennium BC) but different diagenetic environments or storing conditions, to determine the effect of environmental factors on the preservation of DNA. To do so, they carried out the first multiplex approach on ancient specimens using several microsatellites and the sex-determining amelogenin gene, concluding that dryness, low temperature and absence of microorganisms favored DNA preservation.

While during the first years of the field mtDNA was the molecule used because it shows some features which have demonstrated to be adequate in this field, such as being maternally inherited, having higher copy number than nuDNA, lack of recombination and an elevated mutation rate [17], studies using this molecule have major drawbacks [18]. First, mtDNA does not contain information with regard to the masculine evolutionary lineage (that can be different from the feminine one owing to sex-biased demographic processes). In addition, studying just one locus implies having lower resolution in evolutive studies than studying the nuclear genome, as the latter provides data over many thousands of a given individual's ancestors, and not only from one lineage.

Although most aDNA studies rely on a somewhat ideal mitochondrial to nuclear genome ratio of 1.000:1 copies per cell [19, 20], there are no evidences that this ratio can be extrapolated to ancient tissues. In 2009, Schwarz and co-workers [21] compared the number of copies retrieved targeting a 112 bp nuclear amplicon and a similarly long one of the mitochondrial locus using both modern elephant and mammoth samples. The ratios obtained in samples belonging to mammoth surpassed those belonging to modern bone, suggesting preferential mtDNA preservation during diagenesis, being these the first time when this preferential preservation of mtDNA in ancient samples was documented.

Regarding the impact of this new field in the scientific community, it should be acknowledged that during its first years it suffered from an excess of euphoria in relation to the information specialists thought could be obtained. This fact provoked a huge raise in the number of publications that supposedly recovered ancient genetic material from millions of years ago ([22, 23] among others).

Unfortunately, though, things were proven to be not as easy as that [24], and the difficulties to corroborate the authenticity of some of the works made some authors realize the problems inherent to working with aDNA. Consequently, they started proposing some criteria to follow in the methodology of the field and in the results obtained to be regarded as feasible [25, 26].

Later studies postulated that the maximum survival time of DNA was below a million years [27], calling into question studies that claimed to have retrieved older DNA. In fact, almost all these sequences are currently considered artefactual [28], and since then several papers specifying the necessary criteria to endorse the authenticity of these studies have been appearing [29 - 31] to try to guarantee their feasibility.

In spite of the difficulties, the possibilities opened by the emergence of this field were huge. Over the last twenty years this technology has been used for a wide range of purposes, from personal identification to human migrations and to study in depth homo sapiens' (humans') relationship with other hominids.

To begin with, after the first forensic application of DNA fingerprints using highly polymorphic minisatellite loci [32] and the first successful identification of skeletal remains of a murdered child [13], the use of aDNA techniques started to be applied in the forensic context. In addition, in 1995 Alt and co-workers [33] pointed to the necessity of using this information also in archeological contexts, as they considered it essential to confirm or refute hypotheses made by anthropologists about relations in ancient societies.

In brief, the principal fields in which aDNA can provide information are:

Checking for Kinship: In 1997 Hummel and Herrmann [34] followed the path started in [14] providing the second proof of biological kinship at the molecular level for prehistoric individuals in a collective burial. In 2000, a Y chromosomal STR (Y-STR) multiplex adapted to aDNA was presented [35], making it possible the first amplification of Y-STR loci from historical and prehistorical bones of 3,000 years ago. From then on, multiple studies aiming to discern a possible kinship in ancient remains have been carried out ([36, 37] among others), and even used to deep into the relations of Homo neanderthalensis (Neanderthal) population [38].

Characterization and Identification of Historical Figures: This technology has been used to genetically characterize some notorious figures in History and to check the pertinence of some remains to a given individual whenever it was doubtful. For example, to disproof the alleged of a putative son of Marie-Antoniette [39] using mtDNA and to identify the remains of her true son Louis XVII [40], or to corroborate the Romanov family [41] and Nicholas Copernicus [42] remains also using nuDNA information. In other cases like the one of the Italian poet and scholar Francesco Petrarca, the skull and the postcranial skeleton were confirmed to pertain to different individuals, giving credibility to the historically documented profanation of his remains in 1630 [43].

Reconstruction of Human Migrations: In relation to the migrations throughout our species history, aDNA has allowed the accession to the genetic characterization of communities from archeological sites whose antiquity and location is well established, permitting to trace the migrations and the population origins of their main actors [44]. Thus, the accumulating evidences in relation to the relevance of mixture and migration in human history are increasingly coming from genetic analyses of ancient human remains [45, 46], and many of the subjects that aDNA is beginning to shed light on are giving unexpected answers.

For instance, focusing on the origin of the first people that arrived in America, the sequences from two individuals living 24,000 and 17,000 years BP in the Lake Baikal region of Siberia showed that they are phylogenetically closer to current Native Americans than are current Siberians [46].

Ancient DNA studies have also permitted to confirm that agriculture was implemented by one or more populations that differed in their genetic features from local hunter- gatherers [47], and that gene exchange between these two sets of populations, likely during a long period, originated the pattern of variability observed in current Europe [45].

These successes in getting answers about two landmarks in human population migrations set the stage for more studies without the necessity to use inferred data.

Review our Own Species' History: aDNA studies have allowed us the possibility of recovering genetic material from Neanderthals, opening the possibility of knowing which genetic traits were specific of humans.

The first analysis of mtDNA from Neanderthals was published in 1997 [19] grinding up a small sample of bone from the first Neanderthal fossil discovered, in the Neander Valley, Germany. This and a subsequent study [48] showed that the Neanderthal mtDNA showed notorious differences with respect to modern human mtDNA.

A second Neanderthal mtDNA sequence, announced in 2000, from a 29,000 year old Neanderthal found in Mezmaiskaya Cave, Russia [49] proved to be similar to the former one, and also distinct from those of modern humans. This confirms a separation between the Neanderthal and modern human mitochondrial gene pools or with minimum gene flow between them, at least between human males and Neanderthal females.

On the other hand, during the following years it was shown that mtDNA from anatomically modern humans from Europe dating from the same time period as the Neanderthals [50] fit within the range of modern humans, while the Neanderthals remained consistently genetically distinct, showing that modern humans and Neanderthals did not have more genetic similarities during the Pleistocene that were subsequently lost.

Soon afterwards, with the advent of Next-generation sequencing (NGS) techniques in 2005 [51], a quantum leap was taken towards the retrieval of ancient genomes. By avoiding the capillary electrophoresis that limits the throughput of traditional Sanger sequencing, NGS platforms allowed the performance of tens of millions of sequence reactions per machine run, making possible the generation of whole eukaryote genome sequences in a few days [52]. This technology allowed confirming in 2008 that mitogenomes of humans and Neanderthals had not mixed, when Green and co-workers reported the first complete mtDNA Neanderthal sequence from an individual of 38,000 years ago from Vindija Cave, in Croatia [53]. Finally, in 2009 the sequencing of the entire mitochondrial genome of five Neanderthals proved that in spite of the wide geographic area covered by them, their mtDNA genomes were only one third as diverse as modern humans' [54].

These achievements ran in parallel with the accomplishment of the retrieval of nuDNA. Two studies sequenced large amounts of Neanderthal nuDNA and their results were announced in 2006. Green and co-workers announced the sequencing of one million bp of nuDNA of a Neanderthal specimen [55], while the team lead by Noonan sequenced about 65,000 bp from the nuDNA of another specimen [56], setting the stage to study nuDNA from this species. Finally, in 2010 Green's working group obtained a draft sequence of the Neanderthal genome [57] that produced evidences consistent with an interbreeding between anatomically modern humans, and pointed to aspects of the human genome that may have changed since the split between the two species.

Ancient DNA has also permitted to discover that in some places humans' genetic pool still holds the contribution of a yet uncharacterized archaic population that did not belong to the Neanderthal or modern human species and lived in Siberia less tan 50,000 years ago [58]. They were called Denisovans, and since then it has been evidenced that a genetic mixture between a population linked to them and the ancestors of current natives from the Philippines, Australia and New Guinea [58, 59].

Genetically Characterize Extinct Species and Past Infectious Diseases: Apart from the ones of some extinct hominids, other complete mitogenomes of extinct species have been obtained, such as the ones from moa [60], the woolly mammoth [61], from which a draft sequence of the nuclear genome has also been attained [62], the mastodon [63], the extinct cave bear [64] and the woolly rhinoceros [65].

Moreover, ancient genetic material is currently being obtained addressing the causing agents of ancient microbial infections (reviewed by Malgosa and co-workers in 2005 [66] and by Anastasiou and Mitchell in 2013 [67]). Among the most recent successes, it has been possible to shed light on the consequences that the evolution in humans' diet and behavior from the Stone Age to the modern day has had in human populations, using DNA preserved in calcified bacteria on the teeth of ancient skeletons [68]. In addition, it has been possible to suggest that leprosy in America had its origin in Europe by reconstructing entire genomes of Mycobacterium leprae bacteria from five medieval skeletons [69], and to confirm that Yersinia pestis lineages causing the Plague of Justinian and the Black Death were independent emergences from rodents into human beings [70].

7.2. ANCIENT DNA FEATURES AND ENVIRONMENTAL FACTORS

The dynamics of the aged material is a necessary matter of study when aiming to work with aDNA. In brief, the onset of living beings decomposition is driven by a series of reactions which are known altogether as self-digestion or autolysis. With the lack of oxygen suffered by the cells of a given sample, there is an increase in carbon dioxide in the blood, pH drops and residual products pile up poisoning the cells. At the same time, unchecked cellular enzymes, such as lipases or proteases, start disrupting the cells, finally provoking their breakage and the release nutrient-rich fluids [71].

With these cells broken, nutrient-rich fluids are at microorganisms' disposal and putrefaction begins: in the first phase, bacterial enzymes cause the breakdown of tissues by breaking down carbohydrates, lipids and proteins into their single components, generating gases such as methane causing tissue digestion into a fluid consistency, being totally liquefied in the end [72].

In some occasions, this process can be modified either by environmental or endogenous characteristics, which can provoke either the acceleration or the slowdown of the process. In the last case the cadaveric conservation phenomena appear, such as saponification, which inhibits bacterial putrefaction because of adipocere formation due to hydrolysis or hydrogenation from adipose tissues, and mummification, the characteristic ending of a tissue that preserves its integrity when it undergoes desiccation or dehydration. Finally, remains follow one last process named diagenesis, which is basically any chemical, physical, or biological change undergone by organic remains from their initial deposition to their recovery.

7.2.1. Physical and Chemical Agents Damaging DNA

In DNA molecules, the postmortem damage occurs (Fig. **1**) mainly as double-stranded ruptures and oxidative dinucleotide modifications, resulting in the prevention of posterior replications [7, 73], although other changes which also modify the sequence, as for example hydrolytic deamination and depurination, allow the replication to occur and result in variability observed in some of the sequenced clones [19]. Moreover, the action of alkylating agents can also alter the base composition to a variable degree of severity [73] and different types of postmortem damage, such as the miscoding lesions characterized by Pääbo in 1989 [7] can induce that a wrong nucleotide is inserted during the amplification process, leading to amplified sequences that are not a truthful copy of the original. Many works have tried to estimate the real influence of this postmortem damage [74 - 78].

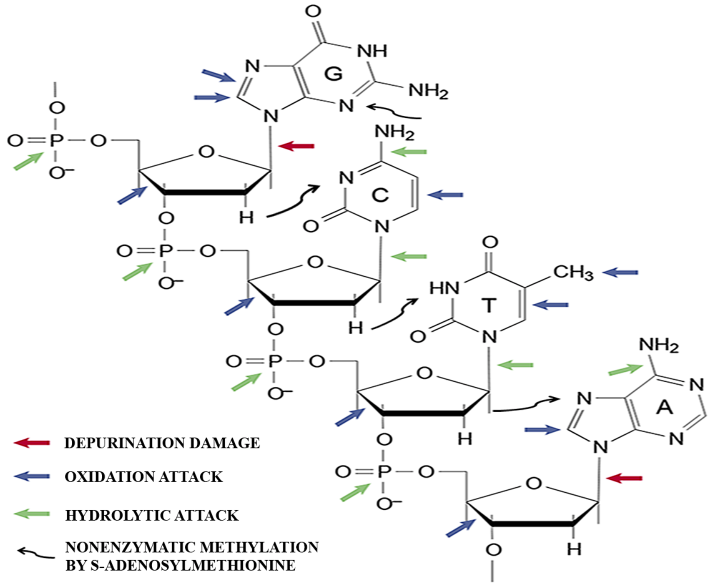


Fig. (1). The principal sites of oxidative damage are marked with blue arrows, the green ones signal the principal points for hydrolytic damage, the red ones point to the principal sites of depurination and the black signal methylation sites (modified from [77]).

Another threat comes from free radicals. They are generated in cells as a by-product of respiration, by ionizing radiation or by the presence of water inside the body, interacting with the DNA molecules and modifying their composition. Moreover, water allows the presence of microorganisms that will use the cellular components as a resource to obtain energy.

Finally, another exogenous factor that is in relation with the environment is the climatology. In general, cold preserved samples will hold retrievable genetic material more frequently than similarly aged ones which have been exposed to warmer temperatures [79], as DNA is best preserved under dry, cool, anaerobic and moderately alkaline environments [80]. In samples exposed to the sun, the ultraviolet radiation (UV) can chemically modify the nitrogenous bases that compose DNA's backbone, specially by the formation of photodimeric lesions [81] or by causing the rupture of the molecule [82].

7.2.2. Main Types of Damage

It has been demonstrated that the most frequent types of damage present in aDNA are two complementary groups of transitions, named "type 1" which encompasses the transitions A→G and T→C, clustered under the expression AT→GC, and "type 2" which covers the CG→TA transitions [76]. All of these transitions result from a single event, the deamination of adenine to hypoxanthine in the first case and of cytosine to uracil in the second one.

In recent studies it has been possible to identify the originally damaged strand, increasing in an 80% the detection of artifacts caused by jumping PCR and providing a way to check how DNA survives after death [78]. The first step in this sense was taken in 2001 when Hofreiter and co-workers [77] showed that the damaging process that would lead from a G to an A analogue change was very unlikely, maybe even impossible. So it could be stated with a big margin of security that any G→A transition seen on the L-strand as a result of molecular damage must have originated as an H-strand C→T change. In contrast, any C→T transition on the L-strand reflects the genuine modification event.

For their part, in 2003 Gilbert and co-workers [78] showed that damage-driven modification of a T→C analogue was also biochemically very improbable. Thus, any L-strand T→C modification will have its origin in an H-strand A→G event, while all the L-strand A→G changes observed will be the result of an original

A→G damage event on that same L-strand (Fig. **2**).

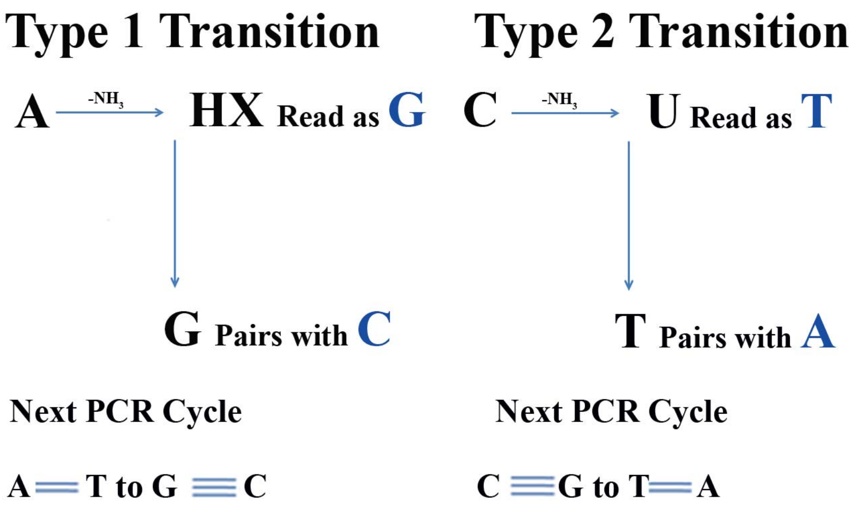


Fig. (2). Type 1 and type 2 transitions caused by molecular damage. Deamination of C→U (read as T) or A→HX (read as G). Modifications appearing on the complementary strand when the damaged bases are added in the next copy process are shown in blue.

To distinguish between postmortem damage and endogenous genetic material the simplest procedure is to compare sequences before and after an enzymatic treatment that reduces the amount of templates carrying postmortem damage and thus the possible misidentification. Different methods employ Uracil-N-Glycosilase (UNG) [7, 77], Endonuclease IV (Endo IV) [7], Alkyladenine DNA glycosylase (AAG) [83], or Pol I polymerase in conjunction with T4 ligase [84].

Among them the most widely used is the UNG treatment, which excises uracil generated as a result of the hydrolytic deamination of cytosine [85]. It diminishes sequence artifacts appearing as a result of this type of postmortem damage that leads to an apparent C→T/G→A mutation [7, 77]. After its application, extracts with different DNA sources will be detected because in them the cloned sequences will still either show the lack of a consensus sequence or some sporadic C→T/G→A base modifications, which now will be recognized as being endogenous to the samples. The main drawback of this treatment is that it will lower the starting-template initial concentration, reducing the amount of samples able to be studied.

7.2.3. Contamination and Its Importance in a DNA Work

Contamination with exogenous material either modern or ancient is the main problem scientists of this area have to deal with. The degradation of the endogenous DNA is translated into higher susceptibility of a sample to get contaminated with exogenous DNA, whether the source is the contact with living tissues or amplified DNA from a previous PCR [86].

It has been suggested that as the more modern contaminant DNA is possibly not as degraded as the endogenous DNA, it is likely to be preferentially amplified when PCR assays are carried out [25]. In addition, some works have demonstrated that the subject under study has become contaminated before its preparation for genetic study (see [30]) and some authors have shown how extremely strict controls can also fail to avoid or detect contamination [74, 75].

Besides this, the fact that the people who handle the samples may hold the same sequences as the ones from some of the studied subjects make them be an easy target for criticism [87], even when additional data as the study of associated fauna or the level of biochemical preservation support the veracity of results [55]. The reasoning behind this is that the most likely source of contamination comes with direct contact, most likely coming from DNA that belongs to the people handling the samples and that permeates throughout dentinal tubules into the pulp cavity (in teeth) and the Haversian system (in bone) [87], even though it probably does not arrive to the osteocytes [88].

Finally, an added problem is that even relatively recent contaminating DNA can mimic sequence damage and allows jumping-PCR between endogenous and contaminant strands [86, 87], increasing the apparent amount of damaged positions in cloned sequences [89].

7.2.4. Reporting Contamination

Thus far contamination happening before the sample entering the laboratory for molecular study has been documented in some studies: many authors have reported modern human DNA in samples where it was not expected to be found, as archeological and historical specimens of calcified tissues from pigs [26], cave bears [77], dogs [89], foxes [90] and Neanderthals [19, 91]. In other cases, this precocious contamination has been detected in human remains through the observation of different DNA haplotypes in cloned sequences pertaining to a single sample or the inconsistency in the results between different samples that belonged to the same skeleton [74, 75, 92]. This kind of contamination in bones and teeth seems to be subject to the level of sample preservation and porosity, with worse preserved samples being more prone to it [87].

The full magnitude of the problem with contamination came to light during the first high-throughput sequencing analyses of Neanderthal samples [55, 56], where it was shown that only a small percentage of the sequences of a Neanderthal bone preserved in a temperate environment belonged to that individual. In addition, just a value of 0.27% was found in a Myotragus specimen, known to be more recent than the aforementioned sample but that had been preserved in a warmer environment in the Balearic Islands [93]. Even in a deep-frozen mammoth bone which apparently showed a good preservation, just 45.4% of the DNA was successfully mapped to the elephant draft genome [20].

Furthermore, recent studies on poorly preserved Neanderthal remains revealed that the majority of the mtDNA had a modern human origin, and many studies have shown that endogenous Neanderthal sequences constitutes no more than 5% of all the sequences obtained [91]. In agreement with this, in 2006 Sampietro and co-workers [86] demonstrated that the predicted contaminants which could most likely have their origin in the people initially washed and cleaned the remains had a higher frequency than the expected from the other participants, supporting previous studies stating that contamination was most susceptible to happen at the initial stage [87, 92]. They also showed that contaminant sequences can suffer miscoding lesion damage after someone has handled them and that the degree of the damage can be quantified, being the level observed among not very old sequences (no more than a decade of age) at levels impossible to distinguish from the ones shown by the sequences considered to be endogenous. This has casted some doubts on a reasoning usually employed to guarantee the veracity of the studies in this field, as is the accumulation of postmortem damage over time.

7.2.5. Decontamination Methodologies

Currently, a universally accepted method to get rid of contamination has not been obtained although several methodologies have been tried ([94, 95] among others). Moreover, most protocols used to avoid contamination have been proven to be unable to eliminate short contaminant DNA fragments at low concentrations [95].

Among the most widely used treatments before an extraction process, it is believed that submersion of bone and tooth samples in diluted bleach serves the purpose of previous handling of ancient remains not influencing the results [96], being also suggested that longer submersion times may minimize the co-extraction of PCR inhibitors [94].

With regard to water, there are discrepancies on whether washing the samples with it can be a source of contamination [86], but some data seem to confirm that it is a critical step in this process [92], and defend the necessity to exclude any sample washing or to do it with sterile water under controlled conditions if totally necessary. In 2008 Bollongino and co-workers [80] stated that it can be the source of non-removable contamination.

The exposition of the samples to UV rays is also problematic because it can also damage the endogenous genetic material, and while breaking the DNA molecules, it is not the elimination but the reduction of the fragment lengths what it finally achieves. So its use continues to be a matter of controversy [94].

On top of that, it has even been proven that individuals preparing reagents and/or making and packaging lab ware pose a particular threat to the possibility of recovering authentic degraded DNA profiles. In fact, some cases of contamination have been found to have occurred this way (see [96]), making it necessary to type all the individuals involved in its preparation. So, whatever the steps taken to avoid it, it should be noted that complete decontamination is not always possible and, thus, results from any aDNA study should be closely scrutinized.

Maintaining high standards within the laboratory is compulsory, and many scientists working with aDNA samples have published papers signaling the points that must be accomplished to obtain reliable results ([29, 31, 87] among others).

CONCLUDING REMARKS

A new field came into play in biological science in the last two decades of the XXth century, namely the study of ancient genetic material. As it could give answers to ancient and new questions, the proverbial virtue of patience was not always regarded in its beginnings. Fortunately though, a general conscience of the difficulties and cautions needed in these studies have been attained over time as this technology's improvement both in terms of resolution and veracity continue to rise.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Pääbo S. Molecular genetic investigations of ancient human remains. Cold Spring Harb Symp Quant Biol 1986; 51(Pt 1): 441-6.
[http://dx.doi.org/10.1101/SQB.1986.051.01.053] [PMID: 3107879]

[2] Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature 1984; 312(5991): 282-4.
[http://dx.doi.org/10.1038/312282a0] [PMID: 6504142]

[3] Pääbo S. Molecular cloning of ancient Egyptian mummy DNA. Nature 1985; 314(6012): 644-5.
[http://dx.doi.org/10.1038/314644a0] [PMID: 3990798]

[4] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 1986; 51(Pt 1): 263-73.
[http://dx.doi.org/10.1101/SQB.1986.051.01.032] [PMID: 3472723]

[5] Mullis KB, Faloona FA. Specific synthesis of DNA *in vitro via* a polymerase-catalyzed chain reaction. Methods Enzymol 1987; 155: 335-50.
[http://dx.doi.org/10.1016/0076-6879(87)55023-6] [PMID: 3431465]

[6] Pääbo S, Wilson AC. Polymerase chain reaction reveals cloning artefacts. Nature 1988; 334(6181): 387-8.
[http://dx.doi.org/10.1038/334387b0] [PMID: 2841606]

[7] Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci USA 1989; 86(6): 1939-43.

[http://dx.doi.org/10.1073/pnas.86.6.1939] [PMID: 2928314]

[8] Hagelberg E, Sykes B, Hedges R. Ancient bone DNA amplified. Nature 1989; 342(6249): 485. [http://dx.doi.org/10.1038/342485a0] [PMID: 2586623]

[9] Thomas RH, Schaffner W, Wilson AC, Pääbo S. DNA phylogeny of the extinct marsupial wolf. Nature 1989; 340(6233): 465-7. [http://dx.doi.org/10.1038/340465a0] [PMID: 2755507]

[10] Horai S, Hayasaka K, Murayama K, Wate N, Koike H, Nakai N. DNA amplification from ancient human skeletal remains and their sequence analysis. P JPN Acad B-Phys 1989; 65: 229-33.

[11] Vargas SR. Material genético de restos óseos humanos. Estudios moleculares en tejidos antiguos. Información Científica y Tecnológica 1989; 11: 19-21.

[12] Hänni C, Laudet V, Sakka M, Bègue A, Stéhelin D. Amplification of mitochondrial DNA fragments from ancient human teeth and bones. C R Acad Sci III 1990; 310(9): 365-70. [PMID: 2113826]

[13] Hagelberg E, Gray IC, Jeffreys AJ. Identification of the skeletal remains of a murder victim by DNA analysis. Nature 1991; 352(6334): 427-9. [http://dx.doi.org/10.1038/352427a0] [PMID: 1861721]

[14] Kurosaki K, Matsushita T, Ueda S. Individual DNA identification from ancient human remains. Am J Hum Genet 1993; 53(3): 638-43. [PMID: 8352274]

[15] Woodward SR, King MJ, Chiu NM, Kuchar MJ, Griggs CW. Amplification of ancient nuclear DNA from teeth and soft tissues. PCR Methods Appl 1994; 3(4): 244-7. [http://dx.doi.org/10.1101/gr.3.4.244] [PMID: 8173514]

[16] Burger J, Hummel S, Hermann B, Henke W. DNA preservation: a microsatellite-DNA study on ancient skeletal remains. Electrophoresis 1999; 20(8): 1722-8. [http://dx.doi.org/10.1002/(SICI)1522-2683(19990101)20:8<1722::AID-ELPS1722>3.0.CO;2-4] [PMID: 10435438]

[17] Ramakrishnan U, Hadly EA. Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies. Mol Ecol 2009; 18(7): 1310-30. [http://dx.doi.org/10.1111/j.1365-294X.2009.04092.x] [PMID: 19281471]

[18] Ballard JW, Whitlock MC. The incomplete natural history of mitochondria. Mol Ecol 2004; 13(4): 729-44. [http://dx.doi.org/10.1046/j.1365-294X.2003.02063.x] [PMID: 15012752]

[19] Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. Neandertal DNA sequences and the origin of modern humans. Cell 1997; 90(1): 19-30. [http://dx.doi.org/10.1016/S0092-8674(00)80310-4] [PMID: 9230299]

[20] Poinar HN, Schwarz C, Qi J, *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 2006; 311(5759): 392-4. [http://dx.doi.org/10.1126/science.1123360] [PMID: 16368896]

[21] Schwarz C, Debruyne R, Kuch M, *et al.* New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic Acids Res 2009; 37(10): 3215-29.

[http://dx.doi.org/10.1093/nar/gkp159] [PMID: 19321502]

[22] Sykes B. Ancient DNA. Less cause for grave concern. Nature 1993; 366(6455): 513. [http://dx.doi.org/10.1038/366513a0] [PMID: 8255287]

[23] Hedges SB, Schweitzer MH. Detecting dinosaur DNA. Science 1995; 268(5214): 1191-2. [http://dx.doi.org/10.1126/science.7761839] [PMID: 7761839]

[24] Zischler H, Höss M, Handt O, von Haeseler A, van der Kuyl AC, Goudsmit J. Detecting dinosaur DNA. Science 1995; 268(5214): 1192-3. [http://dx.doi.org/10.1126/science.7605504] [PMID: 7605504]

[25] Handt O, Höss M, Krings M, Pääbo S. Ancient DNA: methodological challenges. Experientia 1994; 50(6): 524-9. [http://dx.doi.org/10.1007/BF01921720] [PMID: 8020612]

[26] Richards MB, Sykes BC, Hedges RE. Authenticating DNA extracted from ancient skeletal remains. J Archaeol Sci 1995; 22: 291-9. [http://dx.doi.org/10.1006/jasc.1995.0031]

[27] Lindahl T. Recovery of antideluvian DNA. Nature 1993; 365 (700).

[28] Hebsgaard MB, Phillips MJ, Willerslev E. Geologically ancient DNA: fact or artefact? Trends Microbiol 2005; 13(5): 212-20. [http://dx.doi.org/10.1016/j.tim.2005.03.010] [PMID: 15866038]

[29] Cooper A, Poinar HN. Ancient DNA: do it right or not at all. Science 2000; 289(5482): 1139. [http://dx.doi.org/10.1126/science.289.5482.1139b] [PMID: 10970224]

[30] Pääbo S, Poinar H, Serre D, *et al.* Genetic analyses from ancient DNA. Annu Rev Genet 2004; 38: 645-79. [http://dx.doi.org/10.1146/annurev.genet.37.110801.143214] [PMID: 15568989]

[31] Montiel R, Francalacci P, Malgosa A. Ancient DNA and biological anthropology: believers *vs.* skeptics. In: Santos C, Lima M, Eds. Recent advances in molecular biology and evolution: applications to biological anthropology. Trivandrum: Research Signpost 2007; pp. 209-49.

[32] Gill P, Jeffreys AJ, Werrett DJ. Forensic application of DNA fingerprints. Nature 1985; 318(6046): 577-9. [http://dx.doi.org/10.1038/318577a0] [PMID: 3840867]

[33] Alt KW, Vach W, Wahl J. Kinship analysis of skeletal remains from the Bandkeramik mass grave at Talheim, Kreis Heilbronn. Fundberichte aus Baden-Württemberg 1995; 20: 195-217.

[34] Hummel S, Herrmann B. Determination of kinship by aDNA analysis. Anthropol Anz 1997; 55(2): 217-23. [PMID: 9341089]

[35] Schultes T, Hummel S, Herrmann B. Ancient DNA-typing approaches for the determination of kinship in a disturbed collective burial site. Anthropol Anz 2000; 58(1): 37-44. [PMID: 10816784]

[36] Simón M, Jordana X, Armentano N, *et al.* The presence of nuclear families in prehistoric collective burials revisited: the bronze age burial of Montanissell Cave (Spain) in the light of aDNA. Am J Phys

Anthropol 2011; 146(3): 406-13.
[http://dx.doi.org/10.1002/ajpa.21590] [PMID: 21959902]

[37] Deguilloux MF, Pemonge MH, Mendisco F, Thibon D, Cartron I, Castex D. Ancient DNA and kinship analysis of human remains deposited in Merovingian necropolis sarcophagi (Jau Dignac et Loirac, France, 7th–8th century AD). J Archaeol Sci 2014; 42: 373-80.

[38] Lalueza-Fox C, Rosas A, Estalrrich A, *et al.* Genetic evidence for patrilocal mating behavior among Neandertal groups. Proc Natl Acad Sci USA 2011; 108(1): 250-3.
[http://dx.doi.org/10.1073/pnas.1011553108] [PMID: 21173265]

[39] Jehaes E, Decorte R, Peneau A, *et al.* Mitochondrial DNA analysis on remains of a putative son of Louis XVI, King of France and Marie-Antoinette. Eur J Hum Genet 1998; 6(4): 383-95.
[http://dx.doi.org/10.1038/sj.ejhg.5200227] [PMID: 9781047]

[40] Jehaes E, Pfeiffer H, Toprak K, Decorte R, Brinkmann B, Cassiman JJ. Mitochondrial DNA analysis of the putative heart of Louis XVII, son of Louis XVI and Marie-Antoinette. Eur J Hum Genet 2001; 9(3): 185-90.
[http://dx.doi.org/10.1038/sj.ejhg.5200602] [PMID: 11313757]

[41] Gill P, Ivanov PL, Kimpton C, *et al.* Identification of the remains of the Romanov family by DNA analysis. Nat Genet 1994; 6(2): 130-5.
[http://dx.doi.org/10.1038/ng0294-130] [PMID: 8162066]

[42] Bogdanowicz W, Allen M, Branicki W, Lembring M, Gajewska M, Kupiec T. Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus. Proc Natl Acad Sci USA 2009; 106(30): 12279-82.
[http://dx.doi.org/10.1073/pnas.0901848106] [PMID: 19584252]

[43] Caramelli D, Lalueza-Fox C, Capelli C, *et al.* Genetic analysis of the skeletal remains attributed to Francesco Petrarca. Forensic Sci Int 2007; 173(1): 36-40.
[http://dx.doi.org/10.1016/j.forsciint.2007.01.020] [PMID: 17320326]

[44] Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet 2014; 30(9): 377-89.
[http://dx.doi.org/10.1016/j.tig.2014.07.007] [PMID: 25168683]

[45] Skoglund P, Malmström H, Raghavan M, *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 2012; 336(6080): 466-9.
[http://dx.doi.org/10.1126/science.1216304] [PMID: 22539720]

[46] Raghavan M, Skoglund P, Graf KE, *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 2014; 505(7481): 87-91.
[http://dx.doi.org/10.1038/nature12736] [PMID: 24256729]

[47] Brandt G, Haak W, Adler CJ, *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. Science 2013; 342(6155): 257-61.
[http://dx.doi.org/10.1126/science.1241844] [PMID: 24115443]

[48] Krings M, Geisert H, Schmitz RW, Krainitzki H, Pääbo S. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc Natl Acad Sci USA 1999; 96(10): 5581-5.

[http://dx.doi.org/10.1073/pnas.96.10.5581] [PMID: 10318927]

[49] Ovchinnikov IV, Götherström A, Romanova GP, Kharitonov VM, Lidén K, Goodwin W. Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 2000; 404(6777): 490-3.
[http://dx.doi.org/10.1038/35006625] [PMID: 10761915]

[50] Caramelli D, Lalueza-Fox C, Vernesi C, *et al.* Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci USA 2003; 100(11): 6593-7.
[http://dx.doi.org/10.1073/pnas.1130343100] [PMID: 12743370]

[51] Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-80.
[PMID: 16056220]

[52] Kircher M, Kelso J. High-throughput DNA sequencing concepts and limitations. BioEssays 2010; 32(6): 524-36.
[http://dx.doi.org/10.1002/bies.200900181] [PMID: 20486139]

[53] Green RE, Malaspinas AS, Krause J, *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 2008; 134(3): 416-26.
[http://dx.doi.org/10.1016/j.cell.2008.06.021] [PMID: 18692465]

[54] Briggs AW, Good JM, Green RE, *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009; 325(5938): 318-21.
[http://dx.doi.org/10.1126/science.1174462] [PMID: 19608918]

[55] Green RE, Krause J, Ptak SE, *et al.* Analysis of one million base pairs of Neanderthal DNA. Nature 2006; 444(7117): 330-6.
[http://dx.doi.org/10.1038/nature05336] [PMID: 17108958]

[56] Noonan JP, Coop G, Kudaravalli S, *et al.* Sequencing and analysis of Neanderthal genomic DNA. Science 2006; 314(5802): 1113-8.
[http://dx.doi.org/10.1126/science.1131412] [PMID: 17110569]

[57] Green RE, Krause J, Briggs AW, *et al.* A draft sequence of the Neandertal genome. Science 2010; 328(5979): 710-22.
[http://dx.doi.org/10.1126/science.1188021] [PMID: 20448178]

[58] Reich D, Green RE, Kircher M, *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 2010; 468(7327): 1053-60.
[http://dx.doi.org/10.1038/nature09710] [PMID: 21179161]

[59] Meyer M, Kircher M, Gansauge MT, *et al.* A high-coverage genome sequence from an archaic Denisovan individual. Science 2012; 338(6104): 222-6.
[http://dx.doi.org/10.1126/science.1224344] [PMID: 22936568]

[60] Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature 2001; 409(6821): 704-7.
[http://dx.doi.org/10.1038/35055536] [PMID: 11217857]

[61] Krause J, Dear PH, Pollack JL, *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature 2006; 439(7077): 724-7.

[http://dx.doi.org/10.1038/nature04432] [PMID: 16362058]

[62] Miller W, Drautz DI, Ratan A, *et al.* Sequencing the nuclear genome of the extinct woolly mammoth. Nature 2008; 456(7220): 387-90.
[http://dx.doi.org/10.1038/nature07446] [PMID: 19020620]

[63] Rohland N, Malaspinas AS, Pollack JL, Slatkin M, Matheus P, Hofreiter M. Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. PLoS Biol 2007; 5(8): e207.
[http://dx.doi.org/10.1371/journal.pbio.0050207] [PMID: 17676977]

[64] Bon C, Caudy N, de Dieuleveult M, *et al.* Deciphering the complete mitochondrial genome and phylogeny of the extinct cave bear in the Paleolithic painted cave of Chauvet. Proc Natl Acad Sci USA 2008; 105(45): 17447-52.
[http://dx.doi.org/10.1073/pnas.0806143105] [PMID: 18955696]

[65] Willerslev E, Gilbert MT, Binladen J, *et al.* Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. BMC Evol Biol 2009; 9: 95.
[http://dx.doi.org/10.1186/1471-2148-9-95] [PMID: 19432984]

[66] Malgosa A, Montiel R, Díaz N, *et al.* Ancient DNA: a modern look at the infections of the past. In: Pandalai SG, Ed. Recent research developments in microbiology. Trivandrum: Research Signpost 2005; pp. 213-36.

[67] Anastasiou E, Mitchell PD. Palaeopathology and genes: investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. Gene 2013; 528(1): 33-40.
[http://dx.doi.org/10.1016/j.gene.2013.06.017] [PMID: 23792062]

[68] Adler CJ, Dobney K, Weyrich LS, *et al.* Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet 2013; 45(4): 450-455, e1.
[http://dx.doi.org/10.1038/ng.2536] [PMID: 23416520]

[69] Schuenemann VJ, Singh P, Mendum TA, *et al.* Genome-wide comparison of medieval and modern *Mycobacterium leprae.* Science 2013; 341(6142): 179-83.
[http://dx.doi.org/10.1126/science.1238286] [PMID: 23765279]

[70] Wagner DM, Klunk J, Harbeck M, *et al. Yersinia pestis* and the plague of Justinian 541543 AD: a genomic analysis. Lancet Infect Dis 2014; 14(4): 319-26.
[http://dx.doi.org/10.1016/S1473-3099(13)70323-2] [PMID: 24480148]

[71] Vass AA. Beyond the grave: understanding human decomposition. Microbiol Today 2001; 28: 190-2.

[72] Campobasso CP, Di Vella G, Introna F. Factors affecting decomposition and Diptera colonization. Forensic Sci Int 2001; 120(1-2): 18-27.
[http://dx.doi.org/10.1016/S0379-0738(01)00411-X] [PMID: 11457604]

[73] Lindahl T. Instability and decay of the primary structure of DNA. Nature 1993; 362(6422): 709-15.
[http://dx.doi.org/10.1038/362709a0] [PMID: 8469282]

[74] Handt O, Krings M, Ward RH, Pääbo S. The retrieval of ancient human DNA sequences. Am J Hum Genet 1996; 59(2): 368-76.

[PMID: 8755923]

[75] Kolman CJ, Tuross N. Ancient DNA analysis of human populations. Am J Phys Anthropol 2000; 111(1): 5-23.
[http://dx.doi.org/10.1002/(SICI)1096-8644(200001)111:1<5::AID-AJPA2>3.0.CO;2-3] [PMID: 10618586]

[76] Hansen A, Willerslev E, Wiuf C, Mourier T, Arctander P. Statistical evidence for miscoding lesions in ancient DNA templates. Mol Biol Evol 2001; 18(2): 262-5.
[http://dx.doi.org/10.1093/oxfordjournals.molbev.a003800] [PMID: 11158385]

[77] Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet 2001; 2(5): 353-9.
[http://dx.doi.org/10.1038/35072071] [PMID: 11331901]

[78] Gilbert MT, Hansen AJ, Willerslev E, *et al.* Characterization of genetic miscoding lesions caused by postmortem damage. Am J Hum Genet 2003; 72(1): 48-61.
[http://dx.doi.org/10.1086/345379] [PMID: 12489042]

[79] Smith CI, Chamberlain AT, Riley MS, Stringer C, Collins MJ. The thermal history of human fossils and the likelihood of successful DNA amplification. J Hum Evol 2003; 45(3): 203-17.
[http://dx.doi.org/10.1016/S0047-2484(03)00106-4] [PMID: 14580590]

[80] Bollongino R, Tresset A, Vigne JD. Environment and excavation: Pre-lab impacts on ancient DNA analysis. C R Palevol 2008; 7: 91-8.
[http://dx.doi.org/10.1016/j.crpv.2008.02.002]

[81] Besaratinia A, Yoon JI, Schroeder C, Bradforth SE, Cockburn M, Pfeifer GP. Wavelength dependence of ultraviolet radiation-induced DNA damage as determined by laser irradiation suggests that cyclobutane pyrimidine dimers are the principal DNA lesions produced by terrestrial sunlight. FASEB J 2011; 25(9): 3079-91.
[http://dx.doi.org/10.1096/fj.11-187336] [PMID: 21613571]

[82] Slieman TA, Nicholson WL. Artificial and solar UV radiation induces strand breaks and cyclobutane pyrimidine dimers in *Bacillus subtilis* spore DNA. Appl Environ Microbiol 2000; 66(1): 199-205.
[http://dx.doi.org/10.1128/AEM.66.1.199-205.2000] [PMID: 10618224]

[83] Lau AY, Wyatt MD, Glassner BJ, Samson LD, Ellenberger T. Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. Proc Natl Acad Sci USA 2000; 97(25): 13573-8.
[http://dx.doi.org/10.1073/pnas.97.25.13573] [PMID: 11106395]

[84] Pusch CM, Giddings I, Scholz M. Repair of degraded duplex DNA from prehistoric samples using *Escherichia coli* DNA polymerase I and T4 DNA ligase. Nucleic Acids Res 1998; 26(3): 857-9.
[http://dx.doi.org/10.1093/nar/26.3.857] [PMID: 9443981]

[85] Dinner AR, Blackburn GM, Karplus M. Uracil-DNA glycosylase acts by substrate autocatalysis. Nature 2001; 413(6857): 752-5.
[http://dx.doi.org/10.1038/35099587] [PMID: 11607036]

[86] Sampietro ML, Gilbert MT, Lao O, *et al.* Tracking down human contamination in ancient human teeth. Mol Biol Evol 2006; 23(9): 1801-7.
[http://dx.doi.org/10.1093/molbev/msl047] [PMID: 16809622]

[87] Gilbert MT, Bandelt HJ, Hofreiter M, Barnes I. Assessing ancient DNA studies. Trends Ecol Evol (Amst) 2005; 20(10): 541-4.
[http://dx.doi.org/10.1016/j.tree.2005.07.005] [PMID: 16701432]

[88] Malmström H, Storå J, Dalén L, Holmlund G, Götherström A. Extensive human DNA contamination in extracts from ancient dog bones and teeth. Mol Biol Evol 2005; 22(10): 2040-7.
[http://dx.doi.org/10.1093/molbev/msi195] [PMID: 15958782]

[89] Hofreiter M, Betancourt JL, Sbriller AP, Markgraf V, McDonald HG. Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Curá, Neuquén Province, southwest Argentina. Quat Res 2003; 59: 364-78.
[http://dx.doi.org/10.1016/S0033-5894(03)00030-9]

[90] Wandeler P, Smith S, Morin PA, Pettifor RA, Funk SM. Patterns of nuclear DNA degeneration over time case study in historic teeth samples. Mol Ecol 2003; 12(4): 1087-93.
[http://dx.doi.org/10.1046/j.1365-294X.2003.01807.x] [PMID: 12753226]

[91] Serre D, Langaney A, Chech M, *et al.* No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2004; 2(3): E57.
[http://dx.doi.org/10.1371/journal.pbio.0020057] [PMID: 15024415]

[92] Gilbert MT, Hansen AJ, Willerslev E, Turner-Walker G, Collins M. Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA. Int J Osteoarchaeol 2006; 16: 156-64.
[http://dx.doi.org/10.1002/oa.832]

[93] Ramírez O, Gigli E, Bover P, *et al.* Paleogenomics in a temperate environment: shotgun sequencing from an extinct Mediterranean caprine. PLoS One 2009; 4(5): e5670.
[http://dx.doi.org/10.1371/journal.pone.0005670] [PMID: 19461892]

[94] Watt KE. Decontamination techniques in ancient DNA analysis PhD dissertation Burnaby (BC): Simon Fraser University 2005.

[95] Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl EM. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS One 2010; 5(9): e13042.
[http://dx.doi.org/10.1371/journal.pone.0013042] [PMID: 20927390]

[96] Barta JL, Monroe C, Kemp BM. Further evaluation of the efficacy of contamination removal from bone surfaces. Forensic Sci Int 2013; 231(1-3): 340-8.
[http://dx.doi.org/10.1016/j.forsciint.2013.06.004] [PMID: 23890658]

CHAPTER 8

Troubles and Efficiency of aDNA

Marc Simón[*] and **Assumpció Malgosa**

Departament de Biologia Animal, Biologia Vegetal i Ecologia, Unitat d'Antropologia Biològica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona

Abstract: From the establishment of a first set of authenticity criteria to their progressive improvement in parallel with this field's technology advances, the fight to overcome contamination has not ceased over the years. On its part, another problem at the time of recovering ancient genetic material can be caused by properties which may be inherent to the samples, as well as by their interaction with the elements where they are located. A summary from the evolution in these factors' knowledge and the solutions that scientists have given them before the arrival of next-generation sequencing techniques is provided. Finally, a thorough description of the tissues from which ancient genetic material is recovered and the developments to do so from different source organisms is provided.

Keywords: Ancient tissue, Authenticity criteria, Bone, DNA extraction, Feasibility, Hydroxyapatite, Inhibition, Phenol-chloroform, Silica, Sterility, Teeth.

8.1. AUTHENTICITY CRITERIA

As discussed in chapter 7, the difficulty of aDNA studies went hand in hand with the need of guaranteeing the feasibility of the obtained results. In 1989, Pääbo was the first to publish authenticity criteria [1], In which he focused on three points: testing control extracts in parallel with extracts from old specimens; preparing more than one extract from each specimen; and proving the existence of an inverse correlation between amplification efficiency and size of the amplified product. However, serious problems of authenticity persisted: in 1993 Lindahl [2]

[*] **Corresponding author Marc Simon:** Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain; Tel/Fax: +34935811860; E-mail: marcsimon@hotmail.com

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

noted that some claims of very ancient DNA (aDNA) recovery were incompatible with the known biochemical properties of DNA. He recommended the publication of both positive and negative results, their reproducibility, using negative controls and a chemical analysis to check biomolecules' integrity.

Soon after that, Handt and co-workers published in 1994 [3] the first set of integrated criteria a sequence should fulfil to be claimed as ancient, consisting in six points: strict physical separation of the laboratory areas where ancient samples were processed; specially dedicated laboratory clothing to avoid contamination, and accurate cleaning of the work areas with 5% sodium hypochlorite and UV radiation; routine monitoring of contamination; at least two extractions per sample performed at different moments and preferably from different parts of the sample, reporting any incongruent results; consistency with the phylogenetic criterion; and finally, the existence of an inverse relationship between amplification efficiency and molecular length of the amplified fragment. In 1997, Audic and Béraud-Colomb [4] added buffers' sterilization by both autoclave and filtration, and using dedicated pipettes sterilized with UV radiation and aerosol-resistant pipette tips.

However, the accumulation of data evidencing the difficulty to obtain reliable results prompted the need to establish a definitive consensus on which studies could be considered feasible. The proposals went from the establishment of nine comprehensive and very stringent criteria [5], to the approaches suggesting that they should not be set in stone [6 - 8], allowing the investigators to adapt to the particularities of the samples and to apply the logic that each case required.

One of the first studies applying the advice given by Cooper and Poinar [5] was the one by Di Benedetto and co-workers in the same year [9], but the accomplishment of these requisites represented such an effort that the authors concluded that sample sizes for human studies would remain small. Thus, scientists still needed to find a more pragmatic solution, in case it existed.

Further studies showed that while the rigidity proposed for the measures looked correct, they could not always be accomplished following the suggested generic rules [7, 10], but Cooper and Poinar's work had high historical relevance. The main points they suggested to authenticate an aDNA study may be summed up in:

Sterility: all the instrumental required must be sterile, the gloves must be dispensable, a mask, bouffant cap and uniform have to be worn, and physical isolation of the laboratory where treatment and extraction procedures will take place is compulsory. The laboratory should also be positively pressurized, and whenever possible the pieces used should present their full integrity. Finally, it is essential to handle specimens, perform extractions and set up amplifications in dedicated laboratories where no post-PCR work has ever been conducted [11].

Controls of Each Step: all the steps taken must have a control that checks for the absence of contamination. The processes that are carried out before the laboratory work should be monitored and recorded. Also the investigators that had entered in contact with the sample should be genetically characterized. In this sense the first paper to track down the intralaboratory contamination on an aDNA study appeared in 2006 [12], setting a landmark for the future in this kind of studies.

Biochemical Preservation: postmortem changes to tissues cause racemization of the L-amino acids (the only ones incorporated in protein synthesis) resulting in a mixture of L and D enantiomers. In 1996, Poinar and co-workers [13] proved that higher values of D/L in an amino acid known to have a lower degradation rate than others were a sign of contamination by exogenous amino acids.

Appropriate Molecular Behavior: it was first suggested by Pääbo and co-workers in 1988 [14] as a valid authenticity criterion because it was observed that the size of the recovered fragments descended in frequency as the length of the amplicon increased. However, many problems have appeared with regard to the utility of this criterion for distinguishing between endogenous and contaminant DNA. For example, in DNA extracts of the El Sidron Neanderthal contaminant mitochondrial DNA (mtDNA) fragments were found, carrying nucleotide substitutions typical of current humans, but were as short as those from the endogenous Neanderthal mtDNA, while in the Vindija Neanderthal contaminating fragments varied in size [15]. So, it seems that fragment length *per se* is not a reliable estimator of contamination.

Cloning: molecular damage in dead corpses and possible contamination make the cloning of the sequences relevant. Having one sequence each time allows checking if some clones carry the expected kind of lesions in this aged material. Besides, when an ambiguous position in a direct PCR is shown, it helps to clarify which one is the nucleotide originally present and which one is a point mutation due to postmortem damage. Alternatively it can reflect either a true heteroplasmy or the presence of more than one DNA type, indicating contamination.

Reproducibility and Independent Replication: it is advisable to independently replicate the results in both the same and different laboratories, and with both the same and different extracts from the same sample.

Quantitation: target DNA's copy number should be assessed using competitive PCR [16], knowing that if the number of starting templates is low (<1.000), it may be impossible to exclude the possibility of sporadic contamination. Also, the risk of nucleotide misincorporations will be higher if amplifications start from single molecules and DNA sequences are determined from a single amplification [6, 16].

Associated Remains: in human remains' studies where contamination is problematic, survival of DNA similar to the target in associated faunal material is critical supporting evidence, and a good source of human PCR negative controls.

Despite these standard criteria were widely accepted, the debate was not over. In 2004 Pääbo and co-workers [6] remarked again the problem that modern human contamination posed due to its ubiquity. Moreover, they warned about the need of excluding the nuclear insertions of mtDNA and stated that the reproduction in a second laboratory was just warranted when novel or unexpected results were obtained. Finally, they stated that adherence to every criterion in all cases was not warranted as their fulfillment did not guarantee that a DNA sequence was genuinely ancient: there were examples where a specimen was contaminated with a certain DNA sequence and all the criteria had been fulfilled [17, 18]. Thus, the focus should be put on the scientific judgment of the reliability of results, being more of a necessity in aDNA studies than in many other areas of genetics [6].

8.1.1. Phylogenetic and Population Meaning: Refining the Criteria

In 2005 Bandelt [10] added some new criteria to the nine proposed by Cooper and Poinar [5], suggesting that targeted mtDNA fragments of a regional aDNA study should *a priori* be different from potential mtDNA of the archaeological and laboratorial context, as most contamination would not differ from the expected authentic molecule. He also proposed three indicators to be used *a posteriori*: a) the principle of “phylogenetic expectation” [19] by which if the putative ancient mtDNA reflected mtDNA lineages of the human staff handling the samples rather than those expected in the geographic area of the ancient population, contamination would have probably over-run authentic DNA [20]; b) the mosaic structure, which argued that if the putative ancient mtDNA haplotype was composed by fragments fitting with modern mtDNA lineages from different branches of the mtDNA phylogeny, the haplotype would likely constitute some artificial recombinant [21]; and c) the abnormal mutational spectrum, upon which if an agglomeration of unusual mutations was scattered across the mtDNA data, postmortem changes and phantom mutations would have transformed any authentic mtDNA to a degree that resulting sequences would be useless [10]. Lastly, also aiming to increase the feasibility of the works in this field, in 2007 Montiel and co-workers [7] advised against the excessive clonation of the samples, since it might be expensive and generate more problems than it solved due to the increased risk of cross-contamination.

8.1.2. Decontamination: Could it be Achieved?

In 2005, Willerslev and Cooper [22] stated that human contamination in bone and teeth could not be effectively eliminated due to its ubiquity, based on unpublished results by M.T.P. Gilbert, and despite some studies claiming to show that it could be achieved [23]. These two authors [22] also encouraged fulfilling the standard criteria proposed in [5] and proposed some additional ones, like a time-dependent pattern of damage and diversity, decontamination of reagents and specimens, and uracil-N-glycosylase (UNG) treatment to eliminate some postmortem damage.

Soon after that, Gilbert and co-workers [24] published a paper which relaxed to a certain degree the stringency of the criteria, arguing similarly to the previous study by Pääbo and co-workers [6]. They suggested an analysis on a case-by-case basis as to whether the evidences were strong enough to satisfy authenticity criteria, placing the responsibility on authors to self-assess their work in light of the problems inherent to the field. This resulted in the addition of a “tenth commandment" to the original nine points: "Thou shalt interpret the veracity of the data by a critical consideration of all available information" [8].

Finally, the risk of undetected contamination and introduction of specific sequences might be estimated using the contamination events found *a posteriori*. More studies like [12] would provide prior information on this probability.

8.2. INHIBITION AND NATURE OF THE INHIBITORS

Inhibitory substances have also been a matter of concern in this field. Many substances have been identified as inhibitors of the PCR reaction, as they difficult or prevent that the Taq polymerase interacts with the DNA material and carries out the amplification [25]. The interaction of the biological, chemical, and physical taphonomic factors creates variability in the extract composition within and among different burial sites, with different parts of a single piece sometimes bringing very different degrees of amplification success [26]. This stresses the convenience of taking environmental samples when those under study are recovered in the field, to reveal the nature of some potential inhibitors [27] (Fig. **1**). They can have different origins, and ancient samples can display a single inhibitor or a combination of them [28]. Some of the best known ones are the Maillard products, which can block the PCR generating cross-links in the DNA molecule [1]. Also the humic substances, coming from soil or water natural sediments [29] that sometimes give a characteristic color to the extract [30], tannic acid that seems to bind to both DNA and Taq [31], calcium ions, and chelating agents as EDTA. This last substance, though, is useful in many DNA extraction methodologies to demineralize the bone or tooth permitting higher DNA exposure, and to inactivate nucleases that provoke the degradation from the genetic material [32]. So equilibrium must be found using an adequate concentration of this substance in relation to each sample´s characteristics. Finally, there can be residues of porphyrins or their breakdown products, present in many living tissues as blood and that can be present in some cases in archaeological samples [33].

Other PCR inhibitors are endogenous to the biological samples and include calcium ions and collagen from bone [34], blood components as bilirubin [35], saliva, semen, feces [25] or melanin from muscle tissue [36].

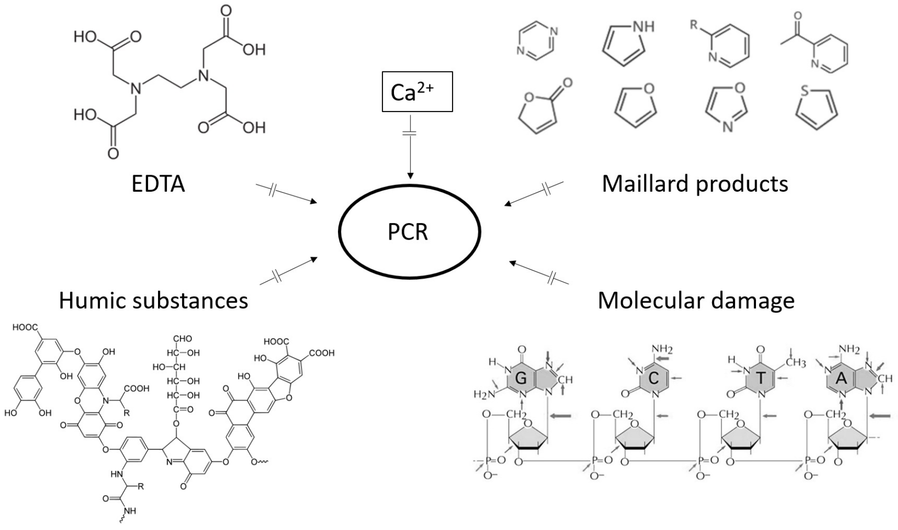


Fig. (1). Some of PCR's main inhibitors (modified from [27]).

8.2.1. Avoiding Inhibition

Some proceedings that eliminate or attenuate the inhibition can be applied before PCR performance [37, 38]. In fact, the widely used phenol–chloroform method allows the recovery of higher DNA quantity [39] but can co-purify with some substances acting as PCR inhibitors [28]. The methods to avoid this problem are divided into two groups: (1) those that remove inhibitors during DNA extraction and purification, and (2) those that diminish the effects of inhibitors by later manipulation of template DNA, PCR reagents, or by incorporating PCR additives.

In relation to inhibitor removing, new protocols enhancing the purity of recovered DNA were developed in the last decade of the XXth century: the silica based purification methods [36, 40]. These methods are based on a silica membrane that virtually binds to every DNA molecule [41] larger than 100 bp but smaller than 10 Kb, excluding the rest of nucleotides, proteins and salts, and seems to eliminate PCR inhibitors and substances giving the brownish color to some bone extracts [42]. However, these methods imply DNA loss, especially of fragments <200 bp, a handicap when working with degraded samples [43], and some of them may also unintentionally incorporate new PCR inhibitors due to reagent carry-over

[44].

Others methods were based on physical characteristics of the inhibitors [28]. The authors claimed that in some cases a pre-PCR cold step helps to overcome the problem leading to the formation of micelles and/or precipitation of the inhibitory substances, leaving DNA free from contamination.

As for the second group, some modifications can be made to increase the chance of a successful amplification. For instance, a DNA sample can be diluted, thereby diluting the present PCR inhibitors [28]. However, aged and badly damaged bones contain limited amounts of DNA template, so this option is often not feasible [42].

In addition, the use of “hot start” Taq DNA polymerase, which requires a pre-PCR heat activation step (95ºC for 11 min), can alleviate the problem of primer-dimers [45]. Hot-start procedures provide other benefits, as room-temperature reagent assembly, increased yield, better specificity and higher efficiency against co-extracted PCR inhibitors [46].

8.3. TISSUES FROM WHERE THE DNA CAN BE EXTRACTED

Different kinds of tissues have been used to obtain aDNA, both from organic and inorganic origin. Teeth have traditionally ranked first among the tissues used as source material in paleogenetic studies, followed by bone material. They are the toughest components of the human body, having the highest resistance to most environmental effects like desiccation and decomposition [47], and their location within the jawbones largely protects them from the environmental and physical conditions that accelerate DNA decay [48]. Besides, their mineral component preserves human DNA by physically excluding microbes and contaminants [49] and enhances its preservation by adsorption to hydroxyapatite (HAP) [50].

With HAP constituting a larger proportion of the overall material in teeth than in bones, the possibilities of successful recovery are usually higher in the former, as confirmed in some works when DNA extracted from teeth was proven to be of higher quality and to yield more quantity ([51, 52] among others).

Teeth are composed by different tissues (enamel, dentine and cementum) and regions (the root(s) and the crown) and DNA is not uniformly distributed among them [53] (Fig. **2a**). The roots, composed of cementum, dentine and pulp, yield more DNA than the crown which is predominantly composed of enamel, an acellular tissue that is the hardest in the skeleton [54].

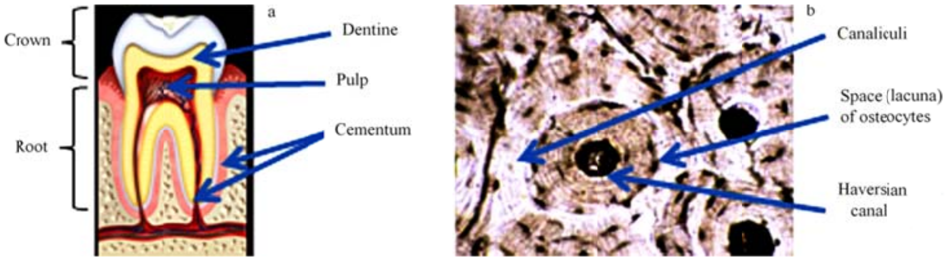


Fig. (2). Inner structure from teeth (**a**), (modified from Matthew Chansky portfolio) and from bone (**b**), (modified from pixshark.com). Tissues from where the DNA is obtained are marked by blue arrows.

The dentine/pulp complex makes up the bulk of the tooth, with pulp being a vascularized connective tissue that contains numerous cell types, providing the richest source of DNA in teeth [55], and dentine lacking nucleated cell bodies but being perforated by tubules containing mtDNA-rich odontoblastic cell processes and nerve fibres [56]. Finally, cementum covers the roots and is made of HAP, collagen and non-collagenous protein, being classified into two types based on the presence or absence of cells (cementocytes, a valuable source of DNA [57]). These facts make pulp and cementum the most valuable sources of nuclear DNA (nuDNA) in the tooth and both these tissues and dentine good sources of mtDNA [55].

On its part, bone shows macroscopically two main architectures. At the jointed ends of long bones and in flat sheet-like bones such as the sternum, it comprises an outer layer of compact bone (cortical bone) that surrounds a load-bearing network of intersecting planes and buttresses called trabeculae (spongy bone), while the mid-shafts of long bones are mainly hollow tubes of cortical bone [58].

Microscopically, bone consists of a hard, apparently homogeneous intercellular material, within or upon which there are various characteristic cell types, as either active or inactive osteoblasts, their osteoprogenitors, osteocytes and osteoclasts.

Osteoblasts are bone cells responsible for bone formation that secrete osteoid, a protein mixture that mineralises with carbonated HAP to become the rigid, load-bearing solid bone mineral [59]; osteocytes originate when osteoblasts become trapped within the matrix they produce, occupying spaces in the bone called lacunae, while osteoclasts carry out bone resorption [60]. These cells are connected inside a network of canals, represented by the Haversian and Volkmann canals, which carries blood vessels, nerves and holds small canaliculi (Fig. **2b**).

Dried, mature bone tissues show this porous structure that can ease the contamination of some extracts [8]. The bone matrix, constituting the majority of the bone, has an inorganic and an organic fraction. The inorganic fraction is made of cryptocrystalline carbonated HAP to which DNA may adsorb [50], and the organic fraction is composed of Type I collagen, whose molecules self aggregate extracellularly into fibrils (fibrillogenesis) that progressively mineralize, and various non-collagenous proteins and glycoproteins [61, 62].

Regarding possible locations of the genetic material, it can survive attached to HAP [50], within bone bioapatite crystals [63] and in intergrown crystal aggregates in fossil bones [64]. A recent study [62] suggested that the HAP component is at least as important as the organic component for DNA recovery, with most ancient samples showing a proportion of mtDNA in each extracted fraction so similar (between 1:4 and 1:1) that discarding either of them would lead to a loss of a relevant proportion of total mtDNA. In 2009 Ottoni and co-workers [65] observed that damaged collagen fibrils do not necessarily relate to poor DNA preservation. Besides, studies using the oxidant NaOCl to clean bones ([20] among others) showed that well preserved DNA can be obtained assuming that intergrown crystal aggregates protect DNA molecules in some fossil bones.

Apart from these two widely used hard tissues, there are at times other kinds of tissues available such as the keratinous tissues like hair and nail [66, 67] as well as horn, feather and scales in vertebrates. These tissues are derived from progenitor cells that undergo cell-death during their biogenesis, so the DNA present is at low level and highly fragmented [68]. As cells undergo natural desiccation during keratinization, DNA survival may be prolonged due to the lack of free water inside them, resulting in a lower number of miscoding lesions [66, 69].

Concerning hair, its metabolically active region is the "root" while the dead, fully keratinized component is the "shaft". Over the past two decades scientists have demonstrated that amplifiable mtDNA and nuDNA can be retrieved from most sources of "fresh" hair [70]. While the hair shaft is composed by dead cells (keratinocytes), recent microscopic imaging have shown nuclear and mitochondrial remnants between macrofibrils within the keratinized hair medial layers forming most of the shaft (cortex), making it probably the most important structure in relation to DNA content [71]. In this sense, studies recovering viable ancient mtDNA from hair cortex have already been published [69].

One of the advantages provided by hair shafts is that they are resistant to contamination from exogenous DNA sources such as bacteria, blood or skin cells, easier to decontaminate than bone and have an apparently lower DNA rate of damage [24, 66, 69]. Gilbert and co-workers [66] proved that hair shafts surpassed comparably stored bone [72] as aDNA source in terms of preservation and concentration of mtDNA relative to nuDNA.

Apparently, mtDNA survives in better condition than nuDNA in hair, as analyses of length distributions of identifiable mtDNA and nuDNA generated through second-generation sequencing confirm [73]. It is likely due to the high energetic demands in hair shafts increasing the number of mitochondria per cell [74].

The advent of "Next-Generation Sequencing" (NGS) [75] has largely overcome the problem regarding the length of the fragments, provoking that hair shafts receive renewed interest as a source for high-throughput sequencing. Recently ancient mitogenomes of taxa including mammoths, humans, thylacines and extinct and extant rhinos [66, 67, 73, 76, 77] have been obtained, culminating in 2010 in the publication of the first high-quality ancient nuclear genome of an approximately 4,500 year old extinct "Saqqaq" Greenlander [78].

Regarding nails, as they share their composition of keratinized cells with hair, they also permit the recovery of mtDNA and nuDNA in fresh samples [79]. Nevertheless, they provide lower yield, at least based on a first comparison [67] made between endogenous mtDNA from century-old toenail samples from black and Javan rhinoceroses and the hair shaft of an ancient woolly rhinoceros.

Preserved ancient fecal matter, or coprolites, can also be a rich source of molecular data [80]. Ancient sequences have been retrieved from coprolites preserved in desert caves and rodent burrows [81], where characteristically dry conditions are thought to be critical for the molecule's long term survival [2]. They can also be a source of keratinous tissues, as many mammals orally groom themselves, their offspring, and conspecifics using the tongue and teeth; also carnivores may intake hair from mammalian preys. In both cases, hair can be excreted in organism's feces, together with a keratinous casing making fecal hairs very resilient at the molecular level [69]. However, coprolites contain a diversity of processed food matter along with the defecator's own sloughed tissue [80]. So a final extract is prone to be a mixture of various sequence templates, making coprolites especially adequate for species-specific shotgun sequencing.

The rest of aDNA samples from animal origin come from tissues that can just be preserved under exceptional conditions [82]. In this sense, the mummification process helps the overall preservation of organic samples and raises the possibility of finding such tissues [83]. It can occur under a wide range of desiccating or anoxic conditions, but only mummification by desiccation can provide high-quality DNA. Notable DNA from human mummified tissues has been recovered, allowing successes such as obtaining the whole genome of the Ötzi mummy, a 5,300 year-old Copper Age individual [84]. However, focusing on ancient Egyptian mummies, there is not a consensus on whether true endogenous DNA can be recovered from them. In 2005 Gilbert and co-workers [85] casted doubts about the authenticity of works dealing with them, predicting that the thermal history of most, if not all, ancient Egyptian material argued against the recovery of endogenous DNA. Despite this, new studies reporting results from this kind of material have continued being published (for example [86]).

Recently, a study that could shed light on this controversy was done applying molecular methodologies to investigate postmortem alterations in a human experimental mummification setting [87]. A lower limb was desiccated in a container of natron salt, simulating the desiccation phase used in ancient Egyptian mummification, and the evolution of skin and muscle tissue was compared taking samples at multiple time points over almost a year. Analyses showed that slight postmortem DNA degradation was present in both tissues, which proved that salt desiccation was an effective method for medium to long-term preservation of human tissues at the molecular level. Now, more studies will have to be done to check the time limits in the efficacy of this method.

As for bacterial DNA, its studies constitute the basis of molecular paleopathology. Until the arrival of techniques using aDNA, the studies from ancient illnesses had been focused on the appearance and prevalence of pathological lesions in human skeletal remains, an approach with many limitations [88], as it was restricted to the individuals with pathognomonic lesions.

Since its inception, this field has been largely driven to the analysis of three organisms, *Yersinia pestis*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* [89 - 91], for which pertinent samples are easily identifiable through historical records or pathological markers. While Mycobacteria are particularly suitable targets for aDNA research due to their protective waxy, hydrophobic and lipid-rich cell wall [92], a recent work [93] has shown that, under the right conditions, sequences can be recovered from organisms with more fragile cell envelopes at least hundreds of years postmortem. For example, despite the difficulty in obtaining ancient treponemal DNA as a result of the bacterium's fragility [94], a recent work has shown that, in particular cases such as neonates with congenital syphilis, it can be recovered if the number of spirochetes in the bones is high enough to allow its preservation [95]. The study of ancient periodontal diseases has also been used as a source to trace back the evolution of those bacteria involved such as caries and *Streptococcus mutans* [96].

In recent years, shotgun metagenomics has been a powerful new tool for paleogenetics to shed light on the emergence, evolution, and spread of microbial pathogens, using contemporaneous, historic and even prehistoric strains [97]. The NGS technologies like microarray-based hybridization capture have permitted to unveil the full bacterial genome of the causative agent of the medieval Black Death epidemic [98] as well as of historical leprosy strains [99]. Furthermore, these techniques are starting to permit the retrieval of ancient viruses, as in the study carried by Appelt and co-workers [100], where they recovered ancient gut viruses from coprolites dating from the Medieval Age, identifying thousands of DNA sequences with homology to known viral families including the still poorly understood bacteriophages, known to play a role in their host's health [101].

Finally, a comparative study done by Willerslev and co-workers in 2004 [102] showed that DNA degradation rates were close to theoretical calculations [17], indicating a limit of approximately 400 thousand years, thus contradicting claims of the obtainment of multi-Ma (million years) DNA sequences or the recovery of putative viable cells of endospores and Proteobacteria from amber and halite [103 - 105] as well as from bacteria from many Ma old permafrost samples [106].

Focusing on DNA from plant remains, many small-scale plant aDNA studies have been done ([107] among others) and recently NGS has been started to be applied ([108] among others). Among the problems that these kinds of samples may show there are the possible presence of polysaccharides, tannins and humic acids, and their large, highly repetitive and heterozygous genomes, confounded by varying ploidy-levels which makes the assemblage of their genomes a challenge [109].

Despite these problems, recent successes include the obtainment of RNA preserved in some ancient seeds, presenting a chance to directly test evolutionary changes in gene expression at a key developmental stage [110], and the genomic characterization of ancient genomes of the plant pathogen *Phytophthora infestans*, the oomycete responsible for the Irish potato famine [111].

CONCLUDING REMARKS

Two things are at the basis of any relevant scientific experiment: empirical proof and the acceptance by the scientific community. This has been especially true in the field of ancient genetic material retrieval. Far from being a handicap, the fact that some notorious names in the field have had at some point different opinions about the feasibility of some renamed studies in this area will result in an overall qualitative improvement, as has been the case in most scientific fields. In this case, clear evidence is given by the vast amount of studies to find out the best source and environmental conditions for the recovery of this material, which have been surely fueled by these debates.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this

publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci USA 1989; 86(6): 1939-43. [http://dx.doi.org/10.1073/pnas.86.6.1939] [PMID: 2928314]

[2] Lindahl T. Recovery of antideluvian DNA. Nature 1993; 365 (700). [http://dx.doi.org/10.1038/365700a0] [PMID: 8413647]

[3] Handt O, Höss M, Krings M, Pääbo S. Ancient DNA: methodological challenges. Experientia 1994; 50(6): 524-9. [http://dx.doi.org/10.1007/BF01921720] [PMID: 8020612]

[4] Audic S, Béraud-Colomb E. Ancient DNA is thirteen years old. Nat Biotechnol 1997; 15(9): 855-8. [http://dx.doi.org/10.1038/nbt0997-855] [PMID: 9306399]

[5] Cooper A, Poinar HN. Ancient DNA: do it right or not at all. Science 2000; 289(5482): 1139. [http://dx.doi.org/10.1126/science.289.5482.1139b] [PMID: 10970224]

[6] Pääbo S, Poinar H, Serre D, *et al.* Genetic analyses from ancient DNA. Annu Rev Genet 2004; 38: 645-79. [http://dx.doi.org/10.1146/annurev.genet.37.110801.143214] [PMID: 15568989]

[7] Montiel R, Francalacci P, Malgosa A. Ancient DNA and biological anthropology: believers *vs.* skeptics. In: Santos C, Lima M, Eds. Recent advances in molecular biology and evolution: Applications to biological anthropology. Trivandrum: Research Signpost 2007; pp. 209-49.

[8] Gilbert MT, Bandelt HJ, Hofreiter M, Barnes I. Assessing ancient DNA studies. Trends Ecol Evol (Amst) 2005; 20(10): 541-4. [http://dx.doi.org/10.1016/j.tree.2005.07.005] [PMID: 16701432]

[9] De Benedetto G, Nasidze IS, Stenico M, *et al.* Mitochondrial DNA sequences in prehistoric human remains from the Alps. Eur J Hum Genet 2000; 8(9): 669-77. [http://dx.doi.org/10.1038/sj.ejhg.5200514] [PMID: 10980572]

[10] Bandelt HJ. Mosaics of ancient mitochondrial DNA: positive indicators of nonauthenticity. Eur J Hum Genet 2005; 13(10): 1106-12. [http://dx.doi.org/10.1038/sj.ejhg.5201476] [PMID: 16077732]

[11] Pääbo S. Amplifying ancient DNA. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, Eds. PCR protocols A guide to methods and applications. New York: Academic Press 1990; pp. 159-66.

[12] Sampietro ML, Gilbert MT, Lao O, *et al.* Tracking down human contamination in ancient human teeth. Mol Biol Evol 2006; 23(9): 1801-7. [http://dx.doi.org/10.1093/molbev/msl047] [PMID: 16809622]

[13] Poinar HN, Höss M, Bada JL, Pääbo S. Amino acid racemization and the preservation of ancient DNA. Science 1996; 272(5263): 864-6.
[http://dx.doi.org/10.1126/science.272.5263.864] [PMID: 8629020]

[14] Pääbo S, Gifford JA, Wilson AC. Mitochondrial DNA sequences from a 7000-year old brain. Nucleic Acids Res 1988; 16(20): 9775-87.
[http://dx.doi.org/10.1093/nar/16.20.9775] [PMID: 3186445]

[15] Briggs AW, Good JM, Green RE, *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009; 325(5938): 318-21.
[http://dx.doi.org/10.1126/science.1174462] [PMID: 19608918]

[16] Handt O, Krings M, Ward RH, Pääbo S. The retrieval of ancient human DNA sequences. Am J Hum Genet 1996; 59(2): 368-76.
[PMID: 8755923]

[17] Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet 2001; 2(5): 353-9.
[http://dx.doi.org/10.1038/35072071] [PMID: 11331901]

[18] Serre D, Langaney A, Chech M, *et al.* No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2004; 2(3): E57.
[http://dx.doi.org/10.1371/journal.pbio.0020057] [PMID: 15024415]

[19] Richards MB, Sykes BC, Hedges RE. Authenticating DNA extracted from ancient skeletal remains. J Archaeol Sci 1995; 22: 291-9.
[http://dx.doi.org/10.1006/jasc.1995.0031]

[20] Kolman CJ, Tuross N. Ancient DNA analysis of human populations. Am J Phys Anthropol 2000; 111(1): 5-23.
[http://dx.doi.org/10.1002/(SICI)1096-8644(200001)111:1<5::AID-AJPA2>3.0.CO;2-3] [PMID: 10618586]

[21] Bandelt HJ. Etruscan artifacts. Am J Hum Genet 2004; 75(5): 919-20.
[http://dx.doi.org/10.1086/425180] [PMID: 15457405]

[22] Willerslev E, Cooper A. Ancient DNA. Proc R Soc B 2005; 272(1558): 3-16.
[http://dx.doi.org/10.1098/rspb.2004.2813] [PMID: 15875564]

[23] Kemp BM, Smith DG. Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. Forensic Sci Int 2005; 154(1): 53-61.
[http://dx.doi.org/10.1016/j.forsciint.2004.11.017] [PMID: 16182949]

[24] Gilbert MT, Hansen AJ, Willerslev E, Turner-Walker G, Collins M. Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA. Int J Osteoarchaeol 2006; 16: 156-64.
[http://dx.doi.org/10.1002/oa.832]

[25] Butler JM, Ed. Advanced topics in forensic DNA typing: methodology. San Diego: Elsevier Academic Press 2012.

[26] Lamers R, Hayter S, Matheson CD. Postmortem miscoding lesions in sequence analysis of human ancient mitochondrial DNA. J Mol Evol 2009; 68(1): 40-55.
[http://dx.doi.org/10.1007/s00239-008-9184-3] [PMID: 19067027]

[27] Simón M, González-Ruiz M, Prats-Muñoz G, Malgosa A. Comparison of two DNA extraction methods in a Spanish Bronze Age burial cave. Quat Int 2012; 247: 358-62. [http://dx.doi.org/10.1016/j.quaint.2011.04.026]

[28] Montiel R, Malgosa A, Subirà E. Overcoming PCR inhibitors in ancient DNA extracts from teeth. Anc Biomol 1997; 1: 221-5.

[29] Watson RJ, Blackwell B. Purification and characterization of a common soil component which inhibits the polymerase chain reaction. Can J Microbiol 2000; 46(7): 633-42. [http://dx.doi.org/10.1139/w00-043] [PMID: 10932357]

[30] Stevenson FJ, Ed. Humus chemistry: genesis, composition, reactions. New York: Wiley Interscience 1982.

[31] Opel KL, Chung D, McCord BR. A study of PCR inhibition mechanisms using real time PCR. J Forensic Sci 2010; 55(1): 25-33. [http://dx.doi.org/10.1111/j.1556-4029.2009.01245.x] [PMID: 20015162]

[32] Loreille OM, Diegoli TM, Irwin JA, Coble MD, Parsons TJ. High efficiency DNA extraction from bone by total demineralization. Forensic Sci Int Genet 2007; 1(2): 191-5. [http://dx.doi.org/10.1016/j.fsigen.2007.02.006] [PMID: 19083754]

[33] Higuchi R. Dr Russ' problem corner. Ancient DNA Newsletter 1992; 1(1): 6-8.

[34] Scholz M, Giddings I, Pusch CM. A polymerase chain reaction inhibitor of ancient hard and soft tissue DNA extracts is determined as human collagen type I. Anal Biochem 1998; 259(2): 283-6. [http://dx.doi.org/10.1006/abio.1998.2676] [PMID: 9618211]

[35] Akane A, Matsubara K, Nakamura H, Takahashi S, Kimura K. Identification of the heme compound copurified with deoxyribonucleic acid (DNA) from bloodstains, a major inhibitor of polymerase chain reaction (PCR) amplification. J Forensic Sci 1994; 39(2): 362-72. [http://dx.doi.org/10.1520/JFS13607J] [PMID: 8195750]

[36] Boom R, Sol CJ, Salimans MM, Jansen CL, Wertheim-van Dillen PM, van der Noordaa J. Rapid and simple method for purification of nucleic acids. J Clin Microbiol 1990; 28(3): 495-503. [PMID: 1691208]

[37] Montiel R, Malgosa A, Francalacci P. Authenticating ancient human mitochondrial DNA. Hum Biol 2001; 73(5): 689-713. [http://dx.doi.org/10.1353/hub.2001.0069] [PMID: 11758690]

[38] Eilert KD, Foran DR. Polymerase resistance to polymerase chain reaction inhibitors in bone. J Forensic Sci 2009; 54(5): 1001-7. [http://dx.doi.org/10.1111/j.1556-4029.2009.01116.x] [PMID: 19686392]

[39] Rohland N, Hofreiter M. Comparison and optimization of ancient DNA extraction. Biotechniques 2007; 42(3): 343-52. [http://dx.doi.org/10.2144/000112383] [PMID: 17390541]

[40] Höss M, Pääbo S. DNA extraction from Pleistocene bones by a silica-based purification method. Nucleic Acids Res 1993; 21(16): 3913-4. [http://dx.doi.org/10.1093/nar/21.16.3913] [PMID: 8396242]

[41] Höss M. More about the silica method. Ancient DNA newsletter 1994; 2(1): 10-2.

[42] Ye J, Ji A, Parra EJ, *et al.* A simple and efficient method for extracting DNA from old and burned bone. J Forensic Sci 2004; 49(4): 754-9.
[http://dx.doi.org/10.1520/JFS2003275] [PMID: 15317190]

[43] Barta JL, Monroe C, Teisberg JE, Winters M. Flanigan, Kemp BM. One of the key characteristics of ancient DNA, low copy number, may be a product of its extraction. J Archaeol Sci 2014; 46: 281-9.
[http://dx.doi.org/10.1016/j.jas.2014.03.030]

[44] Boom R, Sol C, Beld M, Weel J, Goudsmit J, Wertheim-van Dillen P. Improved silica-guanidiniumthiocyanate DNA isolation procedure based on selective binding of bovine alpha-casein to silica particles. J Clin Microbiol 1999; 37(3): 615-9.
[PMID: 9986822]

[45] Birch DE, Kolomodin L, Wang J, *et al.* Simplified hot start PCR. Nature 1996; 381(6581): 445-6.
[http://dx.doi.org/10.1038/381445a0] [PMID: 8632804]

[46] Monroe C, Grier C, Kemp BM. Evaluating the efficacy of various thermo-stable polymerases against co-extracted PCR inhibitors in ancient DNA samples. Forensic Sci Int 2013; 228(1-3): 142-53.
[http://dx.doi.org/10.1016/j.forsciint.2013.02.029] [PMID: 23597751]

[47] Patidar KA, Parwani R, Wanjari S. Effects of high temperature on different restorations in forensic identification: Dental samples and mandible. J Forensic Dent Sci 2010; 2(1): 37-43.
[http://dx.doi.org/10.4103/0974-2948.71056] [PMID: 21189989]

[48] Alvarez García A, Muñoz I, Pestoni C, Lareu MV, Rodríguez-Calvo MS, Carracedo A. Effect of environmental factors on PCR-DNA analysis from dental pulp. Int J Legal Med 1996; 109(3): 125-9.
[http://dx.doi.org/10.1007/BF01369671] [PMID: 8956985]

[49] Milos A, Selmanović A, Smajlović L, *et al.* Success rates of nuclear short tandem repeat typing from different skeletal elements. Croat Med J 2007; 48(4): 486-93.
[PMID: 17696303]

[50] Lindahl T. Instability and decay of the primary structure of DNA. Nature 1993; 362(6422): 709-15.
[http://dx.doi.org/10.1038/362709a0] [PMID: 8469282]

[51] Oota H, Saitou N, Matsushita T, Ueda S. A genetic study of 2,000-year-old human remains from Japan using mitochondrial DNA sequences. Am J Phys Anthropol 1995; 98(2): 133-45.
[http://dx.doi.org/10.1002/ajpa.1330980204] [PMID: 8644875]

[52] Ricaut FX, Fedoseeva A, Keyser-Tracqui C, Crubézy E, Ludes B. Ancient DNA analysis of human neolithic remains found in northeastern Siberia. Am J Phys Anthropol 2005; 126(4): 458-62.
[http://dx.doi.org/10.1002/ajpa.20257] [PMID: 15756672]

[53] Higgins D, Austin JJ. Teeth as a source of DNA for forensic identification of human remains: A review. Sci Justice 2013; 53(4): 433-41.
[http://dx.doi.org/10.1016/j.scijus.2013.06.001] [PMID: 24188345]

[54] Higgins D, Kaidonis J, Austin J, Townsend G, James H, Hughes T. Dentine and cementum as sources of nuclear DNA for use in human identification. Aust J Forensic Sci 2011; 43: 287-95.
[http://dx.doi.org/10.1080/00450618.2011.583278]

[55] Malaver PC, Yunis JJ. Different dental tissues as source of DNA for human identification in forensic cases. Croat Med J 2003; 44(3): 306-9.
[PMID: 12808723]

[56] Mörnstad H, Pfeiffer H, Yoon C, Teivens A. Demonstration and semi-quantification of mtDNA from human dentine and its relation to age. Int J Legal Med 1999; 112(2): 98-100.
[http://dx.doi.org/10.1007/s004140050209] [PMID: 10048666]

[57] Avery JK, Chiego DJ, Eds. Essentials of oral histology and embryology A clinical approach. 3rd ed., Saint Louis: Mosby Elsevier 2006.

[58] Currey JD, Ed. Bones–structure and mechanics. Princeton: Princeton University Press 2002.

[59] Ortner D, Turner-Walker G. The biology of skeletal tissues. In: Ortner D, Ed. Identification of pathological conditions in human skeletal remains. 2nd ed. San Diego: Academic Press 2003; pp. 11-35.
[http://dx.doi.org/10.1016/B978-012528628-2/50039-9]

[60] Nijweide PJ, Burger EH, Feyen JH. Cells of bone: proliferation, differentiation, and hormonal regulation. Physiol Rev 1986; 66(4): 855-86.
[PMID: 3532144]

[61] Tuross N. Recent advances in bone, dentin and enamel biochemistry. Identification of pathological conditions in human skeletal remains. 2nd ed. San Diego: Academic Press 2003; pp. 65-72.
[http://dx.doi.org/10.1016/B978-012528628-2/50042-9]

[62] Campos PF, Craig OE, Turner-Walker G, Peacock E, Willerslev E, Gilbert MT. DNA in ancient bone - where is it located and how should we extract it? Ann Anat 2012; 194(1): 7-16.
[http://dx.doi.org/10.1016/j.aanat.2011.07.003] [PMID: 21855309]

[63] Salamon M, Tuross N, Arensburg B, Weiner S. Relatively well preserved DNA is present in the crystal aggregates of fossil bones. Proc Natl Acad Sci USA 2005; 102(39): 13783-8.
[http://dx.doi.org/10.1073/pnas.0503718102] [PMID: 16162675]

[64] DeNiro MJ, Weiner S. Chemical, enzymatic and spectroscopic characterizaton of “collagen” and other organic fractions from prehistoric bones. Geochim Cosmochim Acta 1988; 52: 2197-206.
[http://dx.doi.org/10.1016/0016-7037(88)90122-6]

[65] Ottoni C, Koon HE, Collins MJ, Penkman KE, Rickards O, Craig OE. Preservation of ancient DNA in thermally damaged archaeological bone. Naturwissenschaften 2009; 96(2): 267-78.
[http://dx.doi.org/10.1007/s00114-008-0478-5] [PMID: 19043689]

[66] Gilbert MT, Tomsho LP, Rendulic S, *et al.* Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 2007; 317(5846): 1927-30.
[http://dx.doi.org/10.1126/science.1146971] [PMID: 17901335]

[67] Willerslev E, Gilbert MT, Binladen J, *et al.* Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. BMC Evol Biol 2009; 9: 95.
[http://dx.doi.org/10.1186/1471-2148-9-95] [PMID: 19432984]

[68] Olsen ME, Bengtsson CF, Bertelsen MF, Willerslev E, Gilbert MT. DNA from keratinous tissue. Part II: feather. Ann Anat 2012; 194(1): 31-5.
[http://dx.doi.org/10.1016/j.aanat.2011.03.003] [PMID: 21489767]

[69] Gilbert MT, Wilson AS, Bunce M, *et al.* Ancient mitochondrial DNA from hair. Curr Biol 2004; 14(12): R463-4.
[http://dx.doi.org/10.1016/j.cub.2004.06.008] [PMID: 15203015]

[70] Baker LE, McCormick WF, Matteson KJ. A silica-based mitochondrial DNA extraction method applied to forensic hair shafts and teeth. J Forensic Sci 2001; 46(1): 126-30.
[http://dx.doi.org/10.1520/JFS14923J] [PMID: 11210897]

[71] Bengtsson CF, Olsen ME, Brandt LØ, *et al.* DNA from keratinous tissue. Part I: hair and nail. Ann Anat 2012; 194(1): 17-25.
[http://dx.doi.org/10.1016/j.aanat.2011.03.013] [PMID: 21530205]

[72] Poinar HN, Schwarz C, Qi J, *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 2006; 311(5759): 392-4.
[http://dx.doi.org/10.1126/science.1123360] [PMID: 16368896]

[73] Gilbert MT, Drautz DI, Lesk AM, *et al.* Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Acad Sci USA 2008; 105(24): 8327-32.
[http://dx.doi.org/10.1073/pnas.0802315105] [PMID: 18541911]

[74] Robin ED, Wong R. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol 1988; 136(3): 507-13.
[http://dx.doi.org/10.1002/jcp.1041360316] [PMID: 3170646]

[75] Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-80.
[http://dx.doi.org/10.1038/nature03959] [PMID: 16056220]

[76] Gilbert MT, Kivisild T, Grønnow B, *et al.* Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. Science 2008; 320(5884): 1787-9.
[http://dx.doi.org/10.1126/science.1159750] [PMID: 18511654]

[77] Miller W, Drautz DI, Janecka JE, *et al.* The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). Genome Res 2009; 19(2): 213-20.
[http://dx.doi.org/10.1101/gr.082628.108] [PMID: 19139089]

[78] Rasmussen M, Li Y, Lindgreen S, *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 2010; 463(7282): 757-62.
[http://dx.doi.org/10.1038/nature08835] [PMID: 20148029]

[79] Cline RE, Laurent NM, Foran DR. The fingernails of Mary Sullivan: developing reliable methods for selectively isolating endogenous and exogenous DNA from evidence. J Forensic Sci 2003; 48(2): 328-33.
[http://dx.doi.org/10.1520/JFS2002107] [PMID: 12664990]

[80] Poinar HN, Hofreiter M, Spaulding WG, *et al.* Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis.* Science 1998; 281(5375): 402-6.
[http://dx.doi.org/10.1126/science.281.5375.402] [PMID: 9665881]

[81] Poinar H, Kuch M, McDonald G, Martin P, Pääbo S. Nuclear gene sequences from a late pleistocene sloth coprolite. Curr Biol 2003; 13(13): 1150-2.
[http://dx.doi.org/10.1016/S0960-9822(03)00450-0] [PMID: 12842016]

[82] Prats-Muñoz G, Galtés I, Armentano N, Cases S, Fernández PL, Malgosa A. Human soft tissue preservation in the Cova des Pas site (Minorca Bronze Age). J Archaeol Sci 2013; 40: 4701-10. [http://dx.doi.org/10.1016/j.jas.2013.07.022]

[83] De Marinis R, Brillante G, Eds. La Mummia del Similaun: Ötzi, l'Uomo Venuto dal Ghiaccio Italy. Venezia: Marsilio 1998.

[84] Keller A, Graefen A, Ball M, *et al.* New insights into the Tyrolean Icemans origin and phenotype as inferred by whole-genome sequencing. Nat Commun 2012; 3: 698. [http://dx.doi.org/10.1038/ncomms1701] [PMID: 22426219]

[85] Gilbert MT, Barnes I, Collins MJ, *et al.* Long-term survival of ancient DNA in Egypt: response to Zink and Nerlich (2003). Am J Phys Anthropol 2005; 128(1): 110-4. [http://dx.doi.org/10.1002/ajpa.20045] [PMID: 15714514]

[86] Khairat R, Ball M, Chang CC, *et al.* First insights into the metagenome of Egyptian mummies using next-generation sequencing. J Appl Genet 2013; 54: 309-25. [http://dx.doi.org/10.1007/s13353-013-0145-1] [PMID: 23553074]

[87] Shved N, Haas C, Papageorgopoulou C, *et al.* Post mortem DNA degradation of human tissue experimentally mummified in salt. PLoS One 2014; 9(10): e110753. [http://dx.doi.org/10.1371/journal.pone.0110753] [PMID: 25337822]

[88] Wood JW, Milner GR, Harpending HC, *et al.* The osteological paradox: Problems of inferring prehistoric health from skeletal samples. Curr Anthropol 1992; 33: 343-70. [http://dx.doi.org/10.1086/204084]

[89] Bos KI, Schuenemann VJ, Golding GB, *et al.* A draft genome of *Yersinia pestis* from victims of the Black Death. Nature 2011; 478(7370): 506-10. [http://dx.doi.org/10.1038/nature10549] [PMID: 21993626]

[90] Suzuki K, Takigawa W, Tanigawa K, *et al.* Detection of *Mycobacterium leprae* DNA from archaeological skeletal remains in Japan using whole genome amplification and polymerase chain reaction. PLoS One 2010; 5(8): e12422. [http://dx.doi.org/10.1371/journal.pone.0012422] [PMID: 20865042]

[91] Donoghue HD. Insights gained from palaeomicrobiology into ancient and modern tuberculosis. Clin Microbiol Infect 2011; 17(6): 821-9. [http://dx.doi.org/10.1111/j.1469-0691.2011.03554.x] [PMID: 21682803]

[92] Donoghue HD, Spigelman M, Greenblatt CL, *et al.* Tuberculosis: from prehistory to Robert Koch, as revealed by ancient DNA. Lancet Infect Dis 2004; 4(9): 584-92. [http://dx.doi.org/10.1016/S1473-3099(04)01133-8] [PMID: 15336226]

[93] Donoghue HD. Insights into ancient leprosy and tuberculosis using metagenomics. Trends Microbiol 2013; 21(9): 448-50. [http://dx.doi.org/10.1016/j.tim.2013.07.007] [PMID: 23932433]

[94] von Hunnius TE, Yang D, Eng B, Waye JS, Saunders SR. Digging deeper into the limits of ancient DNA research on syphilis. J Archaeol Sci 2007; 34(12): 2091-100. [http://dx.doi.org/10.1016/j.jas.2007.02.007]

[95] Montiel R, Solórzano E, Díaz N, *et al.* Neonate human remains: a window of opportunity to the molecular study of ancient syphilis. PLoS One 2012; 7(5): e36371.
[http://dx.doi.org/10.1371/journal.pone.0036371] [PMID: 22567153]

[96] Simón M, Montiel R, Smerling A, *et al.* Molecular analysis of ancient caries. Proc R Soc B 2014; 281(1790) 20140586.
[http://dx.doi.org/10.1098/rspb.2014.0586] [PMID: 25056622]

[97] Whatmore AM. Ancient-pathogen genomics: coming of age? MBio 2014; 5(5): e01676-14.
[http://dx.doi.org/10.1128/mBio.01676-14] [PMID: 25182326]

[98] Bos KI, Stevens P, Nieselt K, Poinar HN, Dewitte SN, Krause J. Yersinia pestis: new evidence for an old infection. PLoS One 2012; 7(11): e49803.
[http://dx.doi.org/10.1371/journal.pone.0049803] [PMID: 23209603]

[99] Schuenemann VJ, Singh P, Mendum TA, *et al.* Genome-wide comparison of medieval and modern *Mycobacterium leprae.* Science 2013; 341(6142): 179-83.
[http://dx.doi.org/10.1126/science.1238286] [PMID: 23765279]

[100] Appelt S, Fancello L, Le Bailly M, Raoult D, Drancourt M, Desnues C. Viruses in a 14th-century coprolite. Appl Environ Microbiol 2014; 80(9): 2648-55.
[http://dx.doi.org/10.1128/AEM.03242-13] [PMID: 24509925]

[101] Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol 2012; 10(9): 607-17.
[http://dx.doi.org/10.1038/nrmicro2853] [PMID: 22864264]

[102] Willerslev E, Hansen AJ, Rønn R, *et al.* Long-term persistence of bacterial DNA. Curr Biol 2004; 14(1): R9-R10.
[http://dx.doi.org/10.1016/j.cub.2003.12.012] [PMID: 14711425]

[103] Cano RJ, Borucki MK. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science 1995; 268(5213): 1060-4.
[http://dx.doi.org/10.1126/science.7538699] [PMID: 7538699]

[104] Vreeland RH, Rosenzweig WD, Powers DW. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature 2000; 407(6806): 897-900.
[http://dx.doi.org/10.1038/35038060] [PMID: 11057666]

[105] Fish SA, Shepherd TJ, McGenity TJ, Grant WD. Recovery of 16S ribosomal RNA gene fragments from ancient halite. Nature 2002; 417(6887): 432-6.
[http://dx.doi.org/10.1038/417432a] [PMID: 12024211]

[106] Shi T, Reeves RH, Gilichinsky DA, Friedmann EI. Characterization of viable bacteria from Siberian permafrost by 16S rDNA sequencing. Microb Ecol 1997; 33(3): 169-79.
[http://dx.doi.org/10.1007/s002489900019] [PMID: 9115180]

[107] Kistler L, Montenegro A, Smith BD, *et al.* Transoceanic drift and the domestication of African bottle gourds in the Americas. Proc Natl Acad Sci USA 2014; 111(8): 2937-41.
[http://dx.doi.org/10.1073/pnas.1318678111] [PMID: 24516122]

[108] Palmer SA, Smith O, Allaby RG. The blossoming of plant archaeogenetics. Special issue: Ancient DNA Ann Anat. 2012; 194: pp. (1)146-56.

[http://dx.doi.org/10.1016/j.aanat.2011.03.012]

[109] Wales N, Andersen K, Cappellini E, Ávila-Arcos MC, Gilbert MT. Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. PLoS One 2014; 9(1): e86827. [http://dx.doi.org/10.1371/journal.pone.0086827] [PMID: 24475182]

[110] Fordyce SL, Ávila-Arcos MC, Rasmussen M, *et al.* Deep sequencing of RNA from ancient maize kernels. PLoS One 2013; 8(1): e50961. [http://dx.doi.org/10.1371/journal.pone.0050961] [PMID: 23326310]

[111] Martin MD, Cappellini E, Samaniego JA, *et al.* Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. Nat Commun 2013; 4: 2172. [http://dx.doi.org/10.1038/ncomms3172] [PMID: 23863894]

CHAPTER 9

aDNA Methodological Revolution

Marc Simón[*] and **Assumpció Malgosa**

Departament de Biologia Animal, Biologia Vegetal i Ecologia, Unitat d'Antropologia Biològica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Abstract: Methods to recover genetic material with the best possible quality have been improving in a similar way to the other areas in this field, culminating with the obtention of the first complete ancient mitogenome in 2001. However, with the arrival of next-generation sequencing in 2005, all these advances can be considered overcome. Using shotgun sequencing as basis and specially-designed microbeads to attach DNA, the advent of the palaeogenomics era has revolutionized the field. Also the enrichment techniques and the growing knowledge of ancient DNA diagenesis add significant achievements, but then again the roof of this technology has yet to be attained.

Keywords: Amelocementary limit, Bleach, DNA extraction, Enrichment, Library preparation, Medullary canal, Microbeads, Mitogenomes, Next-generation sequencing, Pleistocene, Pre-treatment, Priming sites, Purification.

9.1. EVOLUTION OF THE EXTRACTION METHODS

Obtaining DNA from ancient remains is a process that usually needs a pre-treatment of the samples, including removal of the dirt and possible contaminating substances and the solubilization of the genetic material. Although a consensus on which is the best technique has not been reached, bleach is a substance that has been included in the majority of them [1]. Up to now it was thought to be effective, but current studies indicate that this treatment just degrades contaminant DNA, hindering its discernment with respect to the real individual's genetic material [2]. Thus, some authors argue that avoiding to use bleach or any

[*] **Corresponding author Marc Simon:** Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain; Tel/Fax: +34935811860; E-mail: marcsimon@hotmail.com

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)

other substance that is destructive for the DNA molecules is also a viable option, and suggest routine ways of cleaning involving brushing to get rid of some remnants and washing with sterile water instead [3].

Another disagreement appears in the first step of the DNA extraction: how to manage the remains to best perform it. Currently, there are at least three different methodologies which continue to be used: a) powder can be obtained without breaking the piece in an extraction method which does not imply sample destruction, but just its incubation in a specially prepared extraction solution; this method applies to very special museum samples [4]; b) a total breakdown of the piece using liquid nitrogen [5]; c) cutting the piece by its amelocementary limit in the case of teeth and the diaphysis in the case of bones, obtaining powder from the sample with a diamond bur from either the radicular or the medullary canal [6].

To get the best efficiency from the extraction procedures, there are also different strategies. The phenol-chloroform method allows the recovery of higher DNA quantity, but shows some problems like a possible co-purification with substances acting as PCR inhibitors (see “Avoiding inhibition”, chapter II) or the toxicity of the used reagents. The interest in both trying to use new methodologies to avoid the risk for the investigators and in optimizing the genetic material recovery made specialists search for alternatives, leading to the creation of the silica-based purification protocols [7 - 9].

Soon afterwards, the first kits aimed specifically at the aDNA recovery using silica-based spin columns started being used [10]. The method was based on the bind/release of DNA depending on salt concentration. It had the advantage of being faster, using selective binding and being amenable to automation and miniaturization. Since then, more methods using the selective affinity between DNA and silica have appeared, as those based on silica-coated magnetic beads capturing DNA after cell lysis like Promega’s DNA IQ (Promega, USA) or Invitrogen’s Chargeswitch (Invitrogen, USA). Optimizations to make them suitable for robotic systems [11, 12] have also been applied.

Silica gel membranes have also been employed to purify the solutions after having carried out the PCR by pulling out the undesired remaining reagents (for example, the MinElute PCR Purification Kit columns (QIAgen, Netherlands) used by Roeder and co-workers in 2009 [13] or the JetQuick Spin Column Technique (Genomed, Germany) used by Simón and co-workers in 2011 [14]). So far, the best results have been obtained by a process that includes silica spin columns having an additional deposit that enables to load larger volumes. In this process, the binding buffer volume was increased with regard to the extraction buffer and it was composed by guanidine hydrochloride, sodium acetate and isopropanol [15]. Other methods used to purify DNA include filtration using Microcon filter columns [6, 16] or enzymatic hydrolysis using ExoSAP-IT [17].

However, while being safer for the investigators, the problem of inhibition could not be totally overcome with the inception of the silica-based methods [18, 19]. In 2005 though, a novel technology named SCODA ("Synchronous coefficient of drag alteration") (Boreal Genomics, Vancouver) appeared. It improved inhibitor removal with respect to the previous method using silica [20, 21], avoiding the notable loss of nucleic acids due to its automated, minimal-step approach [22]. This method employed alternating electric fields which move DNA to the center of the electrophoretic field whereas the other present substances are moved away from it [20, 23], enabling the obtainment of a good quantity of sample extract. In this way, the DNA recovery from challenging samples such as diluted stains becomes easier.

In order to know if a single extraction method was better than the others many comparative studies have been done over the years, from which none of them has been unanimously concluded to be the best [6, 24 - 26]. However, in 2013 Dabney and co-workers obtained the mitogenome of a Middle Pleistocene cave bear using a modification of the traditional silica extraction method [26, 27], thus making it appear as the best one up until now [15].

9.2. EVOLUTION OF THE SEQUENCING METHODS

To understand the improvement in sequencing technologies one has to look also at the improvement in the amplification methods. Since 1988, when the PCR was first applied to exponentially increase the recovered genetic material from ancient remains [28], until the end of the XXth century, this technique has been the one applied to obtain viable quantities of genetic material from this kind of samples. Possibly, the greatest success attained with the study of material recovered *via* a traditional PCR was the achievement of the first entirely sequenced ancient mitochondrial DNA in 2001, by two different research teams, of the no longer existing moa species [29, 30]. In 2006, Krause and co-workers [31] and Römpler and co-workers [32] enhanced the standard amplification methodology by applying a low amount of multiplex PCRs. Soon some other notorious studies appeared, such as those obtaining the mastodon [33] and cave bear mitogenomes [34] and part of the members of the Romanov family [35].

However, the aim to recover whole genome sequences demanded using new techniques, as the amount of damage present in ancient molecules made traditional approaches unsuitable, because the short length of the authentic molecules could just provide scarce information. The arrival of the so-called "Next-Generation Sequencing" [36] would soon prove to overcome many of the traditional amplification limitations.

9.2.1. Next-Generation Sequencing

The first decade of the XXIst century witnessed the arrival of Next-Generation Sequencing (NGS). It was a fast and cheap way to sequence and analyze large genomes that generally involved amplifying DNA templates and subsequently binding them to a solid surface or to microscopic beads known as microbeads.

The usual NGS techniques required converting DNA segments into DNA libraries as a necessary step previous to sequencing [36 - 39]. The attachment of specially designed DNA segments (adaptors) to both ends of every DNA molecule present accounted for that, providing PCR and sequencing primer ligation sites, and enabling, by parallel sequencing, to obtain millions of template molecules in each reaction [40], reason why the most commonly used term to define this procedure is "massive parallel DNA sequencing". The huge load of obtained sequences caused that almost every region of the genome was well represented. In addition to the high increase in the number of generated sequences, the very low size of the molecules that this technique permits to sequence is another advantageous feature. Moreover, aDNA does not usually need to be chopped before creating the library, which causes the loss of a high amount of molecules, and this is critical in cases when the amount of starting material is diminished in order to accomplish the requirements for the feasibility of the study [41, 42]. Finally, by amplifying all fragments and successfully making them be a part of a library they can be conserved as long as necessary.

The first high-throughput DNA sequencers were the Roche FLX (Roche, Switzerland) and the Illumina GA (Illumina, USA), which permitted the fast sequencing of several different amplified products. Both of them work with double-stranded DNA (dsDNA), and were first created having modern DNA as their main target. They rely on shotgun sequencing, which implies randomly selecting a high amount of DNA molecules among those present in any given sample. Thus, the proportion of any obtained sequence after the application of this technique will be related to the initial one. The first to implement this methodology were 454 Life Sciences, ligating two different adapters to blunt end–repaired dsDNA [36]. Soon afterwards Illumina followed them, using a Y-shaped adapter carrying a T-overhang ligated to both ends of the DNA segments, which had A-overhangs previously added to them [37] (Fig. **1a** and **1b**).

While direct PCR amplification can reconstruct sequences from fragments whose size permits that two PCR primers can hybridize, in library-based techniques DNA fragments which are smaller are fully sequenced due to priming sites used to amplify and sequence them being externally aggregated, and finally the short overlapping sequences are joined into the resulting consensus sequence.

The FLX sequencer's was initially used in the ancient genetic field to carry out the sequencing of the mitochondrial DNA from the Tyrolean Iceman [43], followed by the sequencing of ancient mitochondrial genomes generated from multiplex reactions [44], almost obtaining the whole mitogenomes from more than 30 cave bears. However, at that time these methods continued being expensive and slow, requiring a high amount of PCRs targeting a big number of low size fragments. Therefore, maximizing DNA extraction efficiency was of vital importance [45] and the next step was on the way: enriching genomic regions of interest in the aDNA libraries.

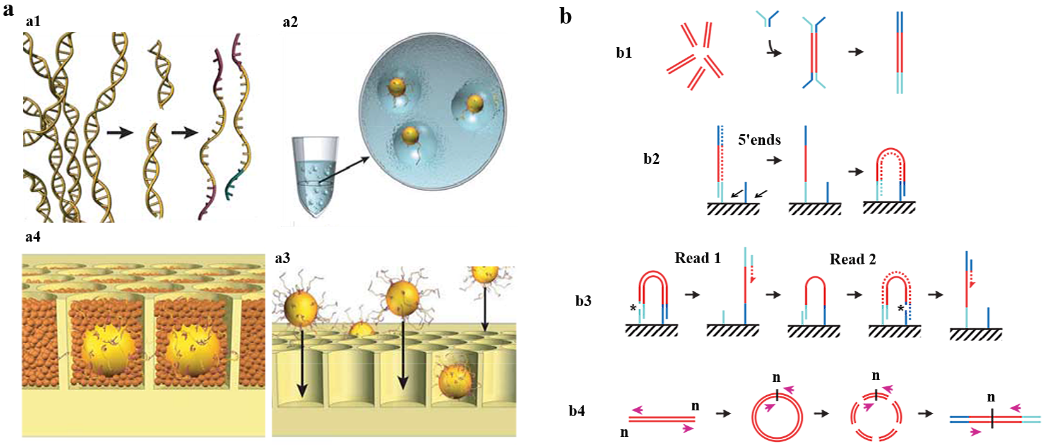


Fig. (1). Methods used by high-throughput DNA sequencers a: 454 Life Sciences method (modified from [36]). b: Illumina Sequencing method (modified from [37]).

Fig. 1a: DNA is isolated, cleaved, joined to adaptors and split into single strands (**a1**). Beads attached in droplets of a PCR-reaction-mixture-in-oil emulsion bind to each fragment; PCR starts inside each droplet from one DNA template (**a2**). DNA is denatured, beads with ssDNA clones are relocated to wells in a fibre-optic slide (**a3**). Smaller beads with enzymes for pyrophosphate sequencing are relocated to each well (**a4**).

Fig. 1b: DNA fragments are created by random chopping, joined to a pair of oligonucleotides and amplified creating ds blunt-ended molecules with distinct adaptors on both ends (**b1**). DNA is denatured and ssDNA anneals to complementary oligonucleotides. A new strand is formed from the original strand, then removed by denaturation. The adaptors' sequence at the 3' end of copied strands anneals to a new surface-bound complementary oligonucleotide, forming a bridge that generates a new site to synthesize a second strand. PCR occurs (**b2**). DNA linearizes by a rupture within one adaptor and is denatured, granting a ss template for read 1. Read 1 products are removed, the template is used to create a bridge, the second strand is re-synthesized and the other strand is cut, granting the template for read 2 (**b3**). In long fragments, their ends are tagged by a biotinylated (B) nucleotide and the fragments circularized. DNA is then cleaved, and the biotinylated junction fragments used as the source material to prepare the samples (**b4**). (◄ : orientation of the reading in relation to the DNA; Y: oligonucleotides; ✕: genomic DNA; ╿: strands generated as a result of cluster formation or sequencing: dotted lines).

9.2.2. Enrichment Techniques

In 2009, Briggs and co-workers [46] proposed a solution to the target-enrichment problem of HTS, published in their analysis of five Neanderthal mitogenomes, named "primer-extension-capture" (PEC). It involved large numbers (>600) of 5′-biotinylated oligonucleotide probes designed to bind specifically to small mitogenomic fragments in HTS libraries prepared using aDNA extracts. Following a series of enzymatic elongation, magnetic streptavidin bead capture and amplification steps, they obtained a significant enrichment of the target in their final library (from 1% to 40% in Neanderthal samples). With such enrichment, libraries from multiple individuals would be able to be sequenced simultaneously by HTS yielding extremely well-covered mitogenomic sequences. The method was able to recover very small targets, as the probes only required a small capture region (approximately 20 bases) and primer-binding sites were not needed, extending the range of material from which aDNA could be retrieved.

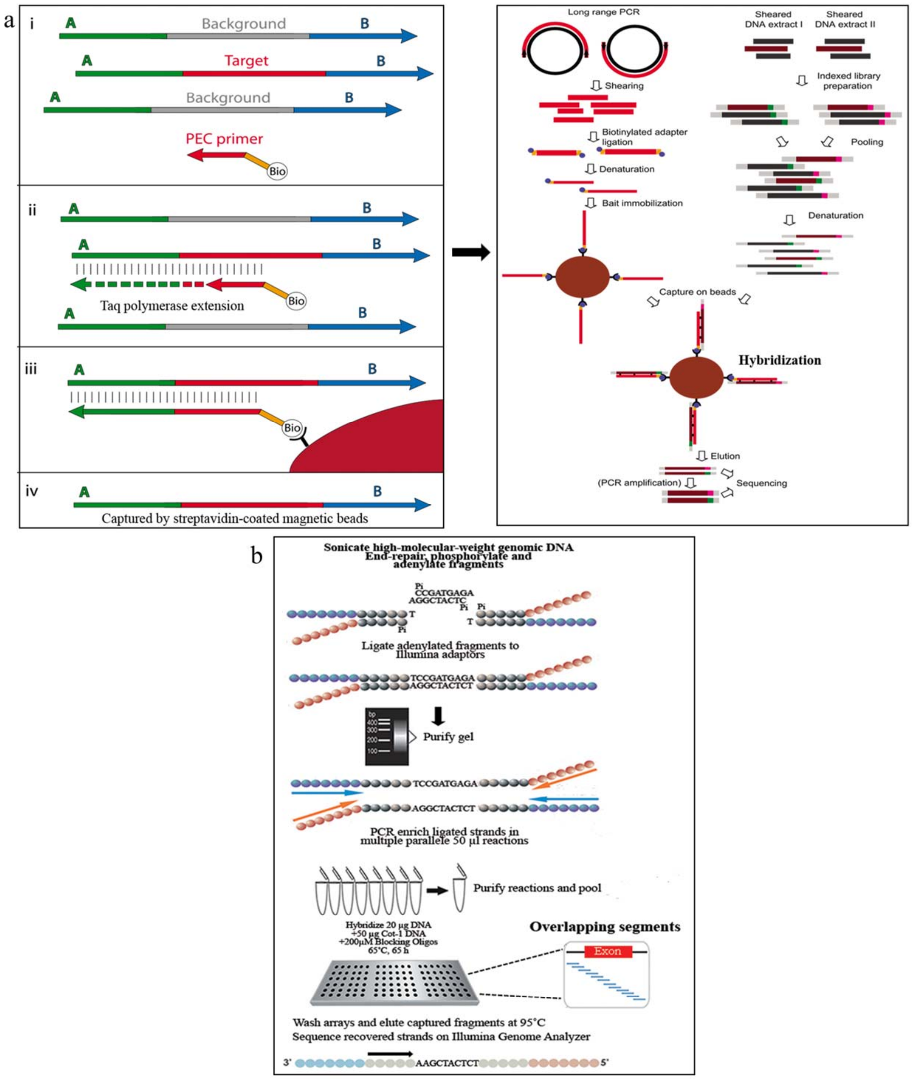


Fig. (2). Enrichment methods. (**a**) Setting (top left) and improvement (top right) of PEC capture methods (modified from [46] and [48]) and (**b**) Microarray capture method (modified from [57]).

This way, the ratio of endogenous *vs.* contaminating hominid DNA could be substantially shifted in favor of the former, obtaining authentic consensus sequences even from strongly contaminated samples when judged by PCR [47]. Other in-solution enrichment methods relied on biotinylated baits to target complementary library inserts, either prepared from modern DNA extracts [48] or designed from known sequences [49]. These methods often delivered complete mitochondrial genome sequences with high depth of coverage [48, 50 - 56]. Finally, another option, named "Microarray-based hybridization capture" [57] (Fig. **2**), also performed well in enriching Neanderthal DNA libraries [58] containing very low endogenous DNA. The use of this enrichment also delivered the full bacterial genome of the causative agent of the medieval Black Death epidemic [59, 60] and of historical leprosy strains [61].

9.2.3. Passing from Second to Third Generation Sequencing Technologies: Advantages of Single-Stranded Library Preparation

In the work that signaled the encounter between second (based on dsDNA and having their representatives in Illumina and FLX sequencers) and third generation sequencing technologies which uses single-stranded DNA (ssDNA) as the starting point, sequencing an actual single strand of DNA and having its representative in the Heliscope Single Molecule Sequencer (Helicos Bioscience, USA), an early Middle Pleistocene horse genome was sequenced at about 1.1-fold coverage combining the former (Illumina) and the latter (Helicos) [53].

Helicos sequencing, based on true single DNA molecule sequencing (tSMS), is a good choice when the purpose is to recover short and damaged DNA segments owing to some of its features. First, as the molecules are initially biotinylated, the subsequent steps happen with DNA bound to streptavidin-coated beads. The only enzymatic treatment required is the poly-adenylation of ssDNA templates, limiting the number of purification steps and avoiding the loss of molecules associated to purification with silica spin columns or carboxylated beads [62], a necessary step when preparing ds libraries. Second, by targeting ssDNA, every single strand break leads to three library templates, in contrast to the single one generated when the target is dsDNA, as the sequencing reaction is primed at any 3'-OH terminus available. In this manner very damaged DNA templates that would be lost in ds libraries are disassembled into multiple fragments in the ss method and every fragment has the possibility to be represented in the library [65] (Fig. **3**). Third, while end modifications on one strand of a ds molecule could prevent that the adaptor ligates when preparing a ds library, when constructing a ss library the opposite strand can still be retrieved. Finally, ds library preparing methods provoke that the original ends of molecules are lost, because they imply blunt-end repair or enzymatic DNA fragmentation [63]. With the ss method, in contrast, the attachment of the adapters does not involve nucleotide removal, enabling the determination of DNA fragmentation patterns in high resolution and potentially providing the possibility to reconstruct each DNA fragment entirely.

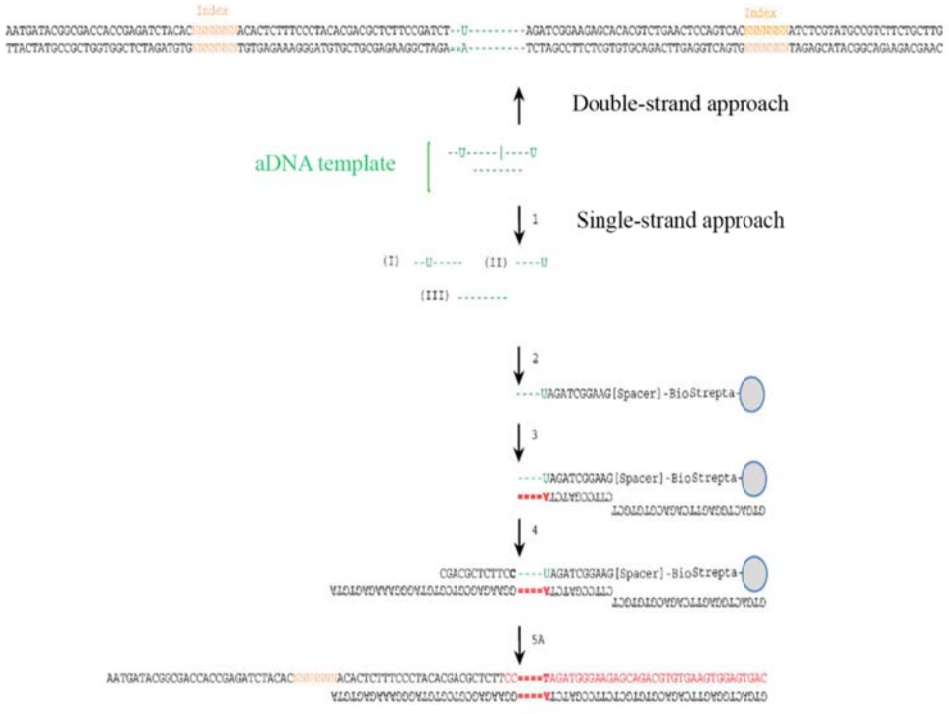


Fig. (3). In contrast to the incorporation of one insert when the library is prepared with dsDNA as a source material, when the starting point is a single DNA strand, three different molecules are generated, so the nicks in the templates of many of the ancient molecules will not prevent that this cleaved molecules get represented in the library. 1: DNA is denatured and three ssDNAs are generated (I, II and III). 2: ss DNA is ligated to a biotinylated adapter. 3: the complementary strand is synthesized, incorporating in this case a G>A misincorporation due to the presence of a uracil residue. 4: DNA is ligated to a new adapter. 5A: start of the PCR amplification, (modified from [65]).

However, a drawback of this methodology is that it only gives the chance to sequence relatively short molecules, as the efficiency of ss ligation drops when the size of the molecule exceeds 120 bp [64]. Most molecules present in old samples do not normally surpass this length, but in samples that have been in an environment enhancing their preservation, such as permafrost, ds library preparation might be more suitable. Moreover, preparing a ss library is more expensive and requires more time than preparing a ds library. So, in case there is a good amount of material that can be used to obtain the extract, using ds libraries may be more suitable even though it means the loss of higher part of the physical integrity of the sample.

Previous studies done with the Heliscope had shown the higher accessibility to this ancient genetic material provided by tSMS [66], suggesting that working with ssDNA would bring some advantages in these studies [66, 67]. In this sense, in 2013 Gansauge and Meyer [68] developed a ssDNA library preparation method specific for the sequencing of aDNA compatible with HTS using Illumina technology. Combining it with the improvement of a silica-based DNA extraction technique [26, 27] to attenuate the bias against ultra-short templates of these libraries, Dabney and co-workers retrieved phylogenetically informative sequences from samples that did not show almost any sequence reaching 50 bp in length, permitting to reconstruct the mitogenome from a Middle Pleistocene cave bear bone (*Ursus deningeri*) found in the Sierra de Atapuerca, Spain [15].

Finally, genetic material of a hominin from the same site and period was obtained in 2014 [69] by Meyer and co-workers. While the first steps of library preparation were the same as the applied in [15], the ubiquity of present-day human contamination prompted them to do an extra step to filter it from endogenous DNA. They combined human-like DNA enrichment with *in silico* selection of the reads holding damages typical of ancient molecules and whose size was characteristic for this kind of material, to obtain the mitogenome of a 400,000 year-old hominin. It constitutes the most recent success in the aDNA field, next to the obtention of the nuclear genome of a 700,000 year-old horse in Canadian permafrost [53].

In both Dabney's and Meyer's studies, targeting ultra-short fragments also reduced the ratio of contaminating molecules, because regions that are 75 bp in length are primarily composed of DNA exogenous to the ancient sample, whereas in shorter fragments the endogenous DNA ratio is much higher [65]. Moreover, the lack of purification steps avoided the reduction in the total amount of DNA recovered that this process implies [62].

To sum up, current methodologies have allowed the sequencing of the oldest mitogenomes from samples not preserved in permafrost to date, namely a 400,000 year-old cave bear and a 400,000 year-old hominin [15, 69], and a 700,000 year-old horse genome [53] from a permafrost-preserved sample, showing that the recovery of genetic material from the Middle Pleistocene is possible.

9.3. RECONSIDERING AUTHENTICITY CRITERIA

As HTS technologies started to be applied [36, 37], it became evident that the authenticity criteria had to be revised. For example, reproducing the results in the same or different laboratories was unsustainable for large-scale DNA sequencing of random molecules due to the time and money required, as well as the fragility of many of the used samples.

The proof that contamination would continue being a problem with this technology came early on, when one of the first works using NGS for aDNA [70] was shown to be affected by larger-sized modern human DNA contaminating fragments and a high sequencing error rate by Wall and Kim [71]. They found that the post-mortem DNA damage, which normally causes cytosine to deaminate and its conversion to uracil, which in turn results in apparent C→T or G→A mutations, made up a significantly lower fraction of the Neanderthal-specific mutations in the work of Green and co-workers [70] than in the one from the same material studied by Noonan and co-workers [72]. It suggested that there should be other causes (apart from postmortem damage suffered by Neanderthal DNA) which were causing the "Neanderthal-specific mutations" in Green´s data [70], pointing towards human contamination. This confirmed that the fact of having constructed the library outside the clean room had been the step when contamination had been introduced, so a new criterion of authenticity should be the building of the library inside it.

NGS techniques brought new knowledge about aDNA diagenesis, which could be used to authenticate endogenous sequences: owing to that, the patterns identified as typical in ancient genetic material have been used to differentiate them from modern molecules. For example, it has been observed that severe degradation modifies aDNA base composition by biasing sequence substitution and fragmentation [73]. Specifically, fragmentation bias has been demonstrated by detecting a relative increase in purines' proportion in the sites that immediately precede the 5' termini in aDNA sequences [74 - 76], While cytosine deamination occurs more frequently in single stranded 5'-overhangs that lead to C- to-T substitutions at 5'- ends [74]. In contrast, complementary G-to-A substitutions at 3'- ends coming with this latter phenomenon are artefacts of the library preparation [47, 67, 74, 77].

More features resulting from next generation library creation are: a) higher GC content in shorter fragments [46, 78, 79] possibly as a result of denaturation of short AT-rich fragments when the library is being prepared [73], b) different polymerases affecting template length and GC content [80], c) bias against fragments starting with a thymine when using AT-overhang ligation protocols [81], d) hairpin loop formation in ssDNA allowing the generation of 3'-end terminal palindromes when creating the library [73] and e) lack of the artefactual complementary G-to-A substitutions at 3'- ends in protocols targeting ssDNA [66, 75, 82]. It is noteworthy that these protocols can provide better yield with respect to background contaminant sequences [66, 68, 75], indicating that the extended protocols for dsDNA have an inherent drawback when creating a library [73].

Finally, two advantages enhancing the recovery of authentic sequences when creating these libraries are: a) the possibility of tagging them in clean labs with project-specific adaptors, allowing the detection of contamination at the library amplification and/or sequencing steps [74], and b) generating such a load of DNA sequences that even rare types of misincorporations are detected.

As NGS techniques have come about almost simultaneously with the chance to genotype extinct species, establishing a consensus about the authenticity criteria in these works has become as relevant as when Cooper and Poinar proposed their nine standard criteria in 2000 [83]. Thus far some suggestions have been:

1. In order to correctly estimate contamination, a two-phase approach has been proposed to sequence ancient genomes, as was done for the Neanderthal mitochondrial genome [77, 78] (Fig. **4**). When good coverage of a genome is finally obtained, the fixed positions showing differences between the extant and the extinct species shall be the key factor to estimate DNA contamination in sequencing of independent libraries from a single individual, such as Briggs and co-workers did in 2009 [46]. To avoid biasing the results, it has to be considered that in these studies small DNA fragments will apparently show less derived alleles in the extinct species than long fragments, as the former are more easily lost in the analysis, overestimating the contamination rates as calculated by Wall and Kim (2007) [71].

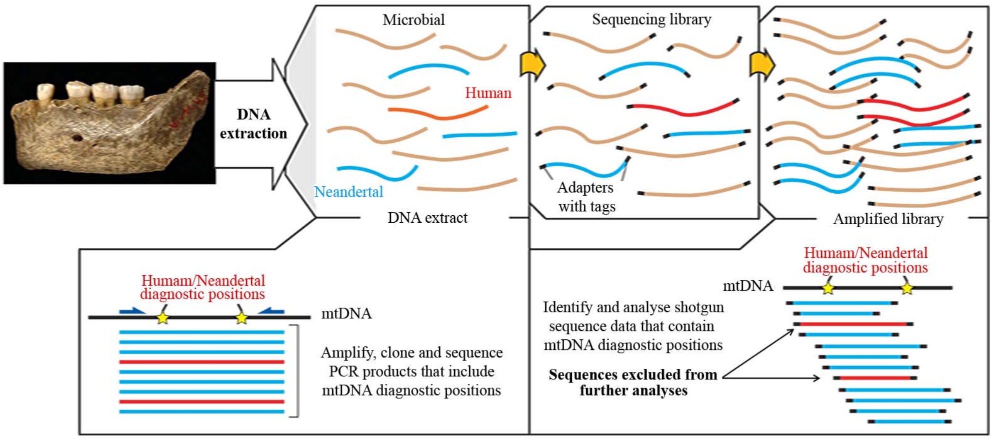

DNA extraction
Microbial
Human
Neandertal
DNA extract
Sequencing library
Adapters with tags
Amplified library
Humam/Neandertal diagnostic positions
mtDNA
Amplify, clone and sequence PCR products that include mtDNA diagnostic positions
Humam/Neandertal diagnostic positions
mtDNA
Identify and analyse shotgun sequence data that contain mtDNA diagnostic positions
Sequences excluded from further analyses

Fig. (4). Amplification and distinction between Human and Neanderthal sequences (modified from [77]).

2. Positions where the sequenced fragment has T or A residues and one of the two species has C or G residues respectively, shall not be considered in the analyses because deaminated C residues in the DNA under study may be their true source [84]. In addition, when working with 454 sequencing, the positions in which two or more contiguous bases exist in one of the species should better be excluded too, as homopolymer length is difficult to score by this method [77].
3. Discounting misincorporations, the number of nucleotide differences between the extinct and the extant species should coincide with the number of reciprocal strand-equivalent differences, where the extinct species’ shows the

complementary nucleotide at the considered site. So this shall be employed for the detection of nucleotide misincorporations. Moreover, reciprocal nucleotide differences should coincide in number provided that the rate and patterns of nucleotide substitutions along both species' evolutionary lineages had a comparable value. So differences in the number between pairs of such reciprocal differences would indicate either that the substitution rate changed in one species (which seems implausible when working with highly related species) or that the observed misincorporations are included within one of the two kinds of differences [85] (Fig. **5**).

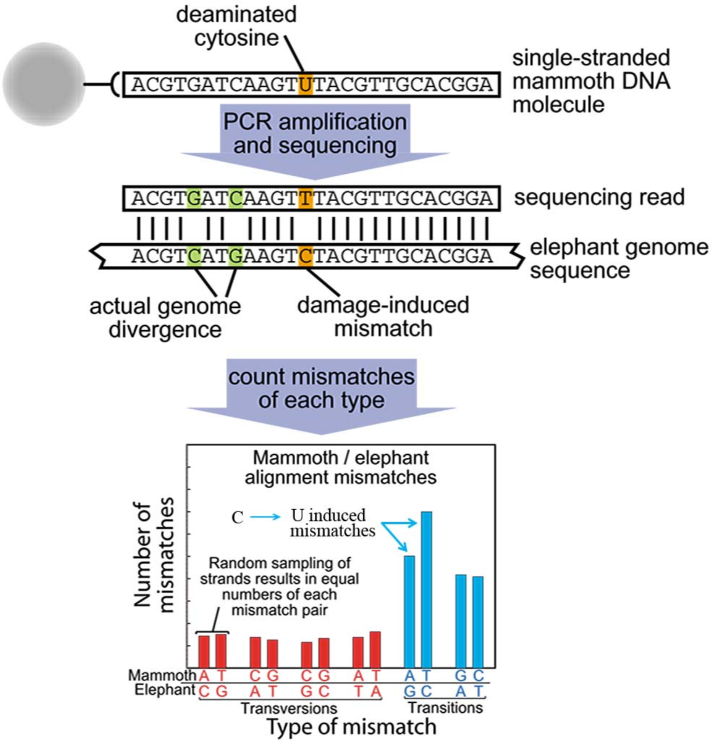

deaminated cytosine
ACGTGATCAAGTUTACGTTGCACGGA
single-stranded mammoth DNA molecule
PCR amplification and sequencing
ACGTGATCAAGTTTACGTTGCACGGA
sequencing read
ACGTCATGAAGTCTACGTTGCACGGA
elephant genome sequence
actual genome divergence
damage-induced mismatch
count mismatches of each type
Mammoth / elephant alignment mismatches
Number of mismatches
C → U induced mismatches
Random sampling of strands results in equal numbers of each mismatch pair
Mammoth A T C G C G A T A T G C
Elephant C G A T G C T A G C A T
Transversions
Transitions
Type of mismatch

Fig. (5). Identification of real differences between mammoths and elephants. Real differences are marked in red, whereas differences owing to nucleotide misincorporations are marked in blue (modified from [85]).

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

As difficult as it is to make predictions in a field where advances continuously surpass the highest expectations, one thing will surely act as a limiting factor to know where the threshold of this field will be set, namely the time limit for the degradation of ancient biomaterial. Recently a study has been published evaluating the absolute limits of DNA survival. This study proposed that in very cold environments DNA dating from more than 1 million years may be recovered [86]. In this sense, the antiquity of the horse mitochondrial genome from Canadian permafrost recovered by Orlando and co-workers [53] opens the possibility of recovering genetic information of Middle Pleistocene animal species, among which Homo Heidelbergensis and Homo erectus would be in the first place in terms of interest. Owing to significant recent advances in biomolecular technology, in the near future it could be possible to recover genetic material from samples where the climate is warmer and DNA degradation advances at a faster pace than in cooler areas [87], maybe providing positive results in the same samples where it had not been possible in the past. Furthermore, knowing that sometimes just a little portion (or none at all) of the DNA from these samples is able to be solubilized into extraction buffer [88] and that many methodologies can just recover a low proportion of the aDNA they hold [89], DNA extraction protocols for ancient samples are likely susceptible to continue being improved.

However, being the move from mitochondrial to nuclear DNA sequencing the next logical step, current techniques would still amplify so many human contaminant DNA that it would be almost impossible to recover whole ancient genomes from the Middle Pleistocene. The *in silico* filtering procedure described in Meyer's work [69] would rule out an excessively high number of sequences, making these studies too costly [65]. This problem can be extrapolated to the study of other species' evolutionary history.

While some methodologies to improve the recovery ratio of endogenous DNA are starting to be applied with moderate success in younger samples as Neanderthals [90, 91] and software to fine tune the selection of ancient molecules is under constant progression (see for instance the enhancement from [67] to [87]), the enrichment of highly damaged DNA templates before the start of the sequencing process is the next logical step to achieve. As with all the previous obstacles, scientists have looked for a way to overcome the problem and successful techniques are starting to be applied in order to enrich the endogenous DNA proportion before library preparation, to once again transform what was thought a chimera into a reality [92], proving that constancy and dedication can pay off in a field that is revolutionizing how we look at the evolutionary history of the live beings.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Kemp BM, Smith DG. Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. Forensic Sci Int 2005; 154(1): 53-61. [http://dx.doi.org/10.1016/j.forsciint.2004.11.017] [PMID: 16182949]

[2] García-Garcerà M, Gigli E, Sanchez-Quinto F, *et al.* Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics. PLoS One 2011; 6(8): e24161. [http://dx.doi.org/10.1371/journal.pone.0024161] [PMID: 21904610]

[3] Higgins D, Austin JJ. Teeth as a source of DNA for forensic identification of human remains: a review. Sci Justice 2013; 53(4): 433-41. [http://dx.doi.org/10.1016/j.scijus.2013.06.001] [PMID: 24188345]

[4] Bolnick DA, Bonine HM, Mata-Míguez J, Kemp BM, Snow MH, LeBlanc SA. Nondestructive sampling of human skeletal remains yields ancient nuclear and mitochondrial DNA. Am J Phys Anthropol 2012; 147(2): 293-300. [http://dx.doi.org/10.1002/ajpa.21647] [PMID: 22183740]

[5] Fernández E, Pérez-Pérez A, Gamba C, *et al.* Ancient DNA analysis of 8000 B.C. near eastern farmers supports an early neolithic pioneer maritime colonization of Mainland Europe through Cyprus and the Aegean Islands. PLoS Genet 2014; 10(6): e1004401. [http://dx.doi.org/10.1371/journal.pgen.1004401] [PMID: 24901650]

[6] Simón M, González-Ruiz M, Prats-Muñoz G, Malgosa A. Comparison of two DNA extraction methods in a Spanish Bronze Age burial cave. Quat Int 2012; 247: 358-62. [http://dx.doi.org/10.1016/j.quaint.2011.04.026]

[7] Boom R, Sol CJ, Salimans MM, Jansen CL, Wertheim-van Dillen PM, van der Noordaa J. Rapid and simple method for purification of nucleic acids. J Clin Microbiol 1990; 28(3): 495-503. [PMID: 1691208]

[8] Höss M, Pääbo S. DNA extraction from Pleistocene bones by a silica-based purification method. Nucleic Acids Res 1993; 21(16): 3913-4.
[http://dx.doi.org/10.1093/nar/21.16.3913] [PMID: 8396242]

[9] Höss M. More about the silica method. Ancient DNA Newsletter 1994; 2(1): 10-2.

[10] Yang DY, Dudar JC, Saunders SR, Waye JS. Removal of PCR inhibitors using silica-based spin columns: application to ancient bones. Can Soc Forensic Sci J 1997; 30(1): 1-5.
[http://dx.doi.org/10.1080/00085030.1997.10757080]

[11] Greenspoon SA, Ban JD, Sykes K, *et al.* Application of the BioMek 2000 laboratory automation workstation and the DNA IQ system to the extraction of forensic casework samples. J Forensic Sci 2004; 49(1): 29-39.
[http://dx.doi.org/10.1520/JFS2003179] [PMID: 14979341]

[12] Frégeau CJ, Lett CM, Fourney RM. Validation of a DNA IQ-based extraction method for TECAN robotic liquid handling workstations for processing casework. Forensic Sci Int Genet 2010; 4(5): 292-304.
[http://dx.doi.org/10.1016/j.fsigen.2009.11.001] [PMID: 20457033]

[13] Roeder AD, Elsmore P, Greenhalgh M, McDonald A. Maximizing DNA profiling success from sub-optimal quantities of DNA: a staged approach. Forensic Sci Int Genet 2009; 3(2): 128-37.
[http://dx.doi.org/10.1016/j.fsigen.2008.12.004] [PMID: 19215883]

[14] Simón M, Jordana X, Armentano N, *et al.* The presence of nuclear families in prehistoric collective burials revisited: the bronze age burial of Montanissell Cave (Spain) in the light of aDNA. Am J Phys Anthropol 2011; 146(3): 406-13.
[http://dx.doi.org/10.1002/ajpa.21590] [PMID: 21959902]

[15] Dabney J, Knapp M, Glocke I, *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA 2013; 110(39): 15758-63.
[http://dx.doi.org/10.1073/pnas.1314445110] [PMID: 24019490]

[16] Forster L, Thomson J, Kutranov S. Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle low copy number (LCN) method for analysis of trace forensic DNA samples. Forensic Sci Int Genet 2008; 2(4): 318-28.
[http://dx.doi.org/10.1016/j.fsigen.2008.04.005] [PMID: 19083842]

[17] Smith PJ, Ballantyne J. Simplified low-copy-number DNA analysis by post-PCR purification. J Forensic Sci 2007; 52(4): 820-9.
[http://dx.doi.org/10.1111/j.1556-4029.2007.00470.x] [PMID: 17553095]

[18] Kemp BM, Monroe C, Smith DG. Repeat silica extraction: a simple technique for the removal of PCR inhibitors from DNA extracts. J Archaeol Sci 2006; 33: 1680-9.
[http://dx.doi.org/10.1016/j.jas.2006.02.015]

[19] Lee HY, Park MJ, Kim NY, Sim JE, Yang WI, Shin KJ. Simple and highly effective DNA extraction methods from old skeletal remains using silica columns. Forensic Sci Int Genet 2010; 4(5): 275-80.
[http://dx.doi.org/10.1016/j.fsigen.2009.10.014] [PMID: 20457067]

[20] Marziali A, Pel J, Bizzotto D, Whitehead LA. Novel electrophoresis mechanism based on synchronous

alternating drag perturbation. Electrophoresis 2005; 26(1): 82-90.
[http://dx.doi.org/10.1002/elps.200406140] [PMID: 15624147]

[21] Broemeling DJ, Pel J, Gunn DC, *et al.* An instrument for automated purification of nucleic acids from contaminated forensic samples. JALA Charlottesv Va 2008; 13(1): 40-8.
[http://dx.doi.org/10.1016/j.jala.2007.10.008] [PMID: 18438455]

[22] Schmedes S, Marshall P, King JL, Budowle B. Effective removal of co-purified inhibitors from extracted DNA samples using synchronous coefficient of drag alteration (SCODA) technology. Int J Legal Med 2013; 127(4): 749-55.
[http://dx.doi.org/10.1007/s00414-012-0810-7] [PMID: 23254459]

[23] Pel J, Broemeling D, Mai L, *et al.* Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. Proc Natl Acad Sci USA 2009; 106(35): 14796-801.
[http://dx.doi.org/10.1073/pnas.0907402106] [PMID: 19706437]

[24] Yang DY, Eng B, Waye JS, Dudar JC, Saunders SR. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am J Phys Anthropol 1998; 105(4): 539-43.
[http://dx.doi.org/10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1] [PMID: 9584894]

[25] Bouwman AS, Brown TA. Comparison between silica-based methods for the extraction of DNA from human bones from 18th–mid-19th century London. Anc Biomol 2002; 4: 173-8.
[http://dx.doi.org/10.1080/1358612021000028470]

[26] Rohland N, Hofreiter M. Comparison and optimization of ancient DNA extraction. Biotechniques 2007; 42(3): 343-52.
[http://dx.doi.org/10.2144/000112383] [PMID: 17390541]

[27] Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. Nat Protoc 2007; 2(7): 1756-62.
[http://dx.doi.org/10.1038/nprot.2007.247] [PMID: 17641642]

[28] Pääbo S, Wilson AC. Polymerase chain reaction reveals cloning artefacts. Nature 1988; 334(6181): 387-8.
[http://dx.doi.org/10.1038/334387b0] [PMID: 2841606]

[29] Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature 2001; 409(6821): 704-7.
[http://dx.doi.org/10.1038/35055536] [PMID: 11217857]

[30] Haddrath O, Baker AJ. Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. Proc Biol Sci 2001; 268(1470): 939-45.
[http://dx.doi.org/10.1098/rspb.2001.1587] [PMID: 11370967]

[31] Krause J, Dear PH, Pollack JL, *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature 2006; 439(7077): 724-7.
[http://dx.doi.org/10.1038/nature04432] [PMID: 16362058]

[32] Römpler H, Dear PH, Krause J, *et al.* Multiplex amplification of ancient DNA. Nat Protoc 2006; 1(2): 720-8.
[http://dx.doi.org/10.1038/nprot.2006.84] [PMID: 17406302]

[33] Rohland N, Malaspinas AS, Pollack JL, Slatkin M, Matheus P, Hofreiter M. Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. PLoS Biol 2007; 5(8): e207.
[http://dx.doi.org/10.1371/journal.pbio.0050207] [PMID: 17676977]

[34] Krause J, Unger T, Noçon A, *et al.* Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol 2008; 8: 220.
[http://dx.doi.org/10.1186/1471-2148-8-220] [PMID: 18662376]

[35] Rogaev EI, Grigorenko AP, Moliaka YK, *et al.* Genomic identification in the historical case of the Nicholas II royal family. Proc Natl Acad Sci USA 2009; 106(13): 5258-63.
[http://dx.doi.org/10.1073/pnas.0811190106] [PMID: 19251637]

[36] Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; 437(7057): 376-80.
[PMID: 16056220]

[37] Bentley DR, Balasubramanian S, Swerdlow HP, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456(7218): 53-9.
[http://dx.doi.org/10.1038/nature07517] [PMID: 18987734]

[38] McKernan KJ, Peckham HE, Costa GL, *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 2009; 19(9): 1527-41.
[http://dx.doi.org/10.1101/gr.091868.109] [PMID: 19546169]

[39] Rothberg JM, Hinz W, Rearick TM, *et al.* An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011; 475(7356): 348-52.
[http://dx.doi.org/10.1038/nature10242] [PMID: 21776081]

[40] Ho SY, Gilbert MT. Ancient mitogenomics. Mitochondrion 2010; 10(1): 1-11.
[http://dx.doi.org/10.1016/j.mito.2009.09.005] [PMID: 19788938]

[41] Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nat Protoc 2008; 3(2): 267-78.
[http://dx.doi.org/10.1038/nprot.2007.520] [PMID: 18274529]

[42] Maricic T, Pääbo S. Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. Biotechniques 2009; 46(1): 51-52, 54-57.
[http://dx.doi.org/10.2144/000113042] [PMID: 19301622]

[43] Ermini L, Olivieri C, Rizzi E, *et al.* Complete mitochondrial genome sequence of the Tyrolean Iceman. Curr Biol 2008; 18(21): 1687-93.
[http://dx.doi.org/10.1016/j.cub.2008.09.028] [PMID: 18976917]

[44] Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M. Direct multiplex sequencing (DMPS)a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Res 2009; 19(10): 1843-8.
[http://dx.doi.org/10.1101/gr.095760.109] [PMID: 19635845]

[45] Hofreiter M, Paijmans JL, Goodchild H, *et al.* The future of ancient DNA: Technical advances and conceptual shifts. BioEssays 2015; 37(3): 284-93.

[http://dx.doi.org/10.1002/bies.201400160] [PMID: 25413709]

[46] Briggs AW, Good JM, Green RE, *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 2009; 325(5938): 318-21.
[http://dx.doi.org/10.1126/science.1174462] [PMID: 19608918]

[47] Krause J, Briggs AW, Kircher M, *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol 2010; 20(3): 231-6.
[http://dx.doi.org/10.1016/j.cub.2009.11.068] [PMID: 20045327]

[48] Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 2010; 5(11): e14004.
[http://dx.doi.org/10.1371/journal.pone.0014004] [PMID: 21103372]

[49] Ávila-Arcos MC, Cappellini E, Romero-Navarro JA, *et al.* Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Sci Rep 2011; 1(74): 74.
[PMID: 22355593]

[50] Krause J, Fu Q, Good JM, *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 2010; 464(7290): 894-7.
[http://dx.doi.org/10.1038/nature08976] [PMID: 20336068]

[51] Horn S, Durka W, Wolf R, *et al.* Mitochondrial genomes reveal slow rates of molecular evolution and the timing of speciation in beavers (Castor), one of the largest rodent species. PLoS One 2011; 6(1): e14622.
[http://dx.doi.org/10.1371/journal.pone.0014622] [PMID: 21307956]

[52] Fu Q, Mittnik A, Johnson PL, *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 2013; 23(7): 553-9.
[http://dx.doi.org/10.1016/j.cub.2013.02.044] [PMID: 23523248]

[53] Orlando L, Ginolhac A, Zhang G, *et al.* Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. Nature 2013; 499(7456): 74-8.
[http://dx.doi.org/10.1038/nature12323] [PMID: 23803765]

[54] Thalmann O, Shapiro B, Cui P, *et al.* Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. Science 2013; 342(6160): 871-4.
[http://dx.doi.org/10.1126/science.1243650] [PMID: 24233726]

[55] Vilstrup JT, Seguin-Orlando A, Stiller M, *et al.* Mitochondrial phylogenomics of modern and ancient equids. PLoS One 2013; 8(2): e55950.
[http://dx.doi.org/10.1371/journal.pone.0055950] [PMID: 23437078]

[56] Zhang H, Paijmans JL, Chang F, *et al.* Morphological and genetic evidence for early Holocene cattle management in northeastern China. Nat Commun 2013; 4: 2755.
[PMID: 24202175]

[57] Hodges E, Smith AD, Kendall J, *et al.* High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res 2009; 19(9): 1593-605.
[http://dx.doi.org/10.1101/gr.095190.109] [PMID: 19581485]

[58] Burbano HA, Hodges E, Green RE, *et al.* Targeted investigation of the Neandertal genome by array-based sequence capture. Science 2010; 328(5979): 723-5.

[http://dx.doi.org/10.1126/science.1188046] [PMID: 20448179]

[59] Schuenemann VJ, Bos K, DeWitte S, *et al.* Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. Proc Natl Acad Sci USA 2011; 108(38): E746-52.
[http://dx.doi.org/10.1073/pnas.1105107108] [PMID: 21876176]

[60] Bos KI, Stevens P, Nieselt K, Poinar HN, Dewitte SN, Krause J. Yersinia pestis: new evidence for an old infection. PLoS One 2012; 7(11): e49803.
[http://dx.doi.org/10.1371/journal.pone.0049803] [PMID: 23209603]

[61] Schuenemann VJ, Singh P, Mendum TA, *et al.* Genome-wide comparison of medieval and modern Mycobacterium leprae. Science 2013; 341(6142): 179-83.
[http://dx.doi.org/10.1126/science.1238286] [PMID: 23765279]

[62] DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res 1995; 23(22): 4742-3.
[http://dx.doi.org/10.1093/nar/23.22.4742] [PMID: 8524672]

[63] Adey A, Morrison HG, Asan , *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. Genome Biol 2010; 11(12): R119.
[http://dx.doi.org/10.1186/gb-2010-11-12-r119] [PMID: 21143862]

[64] Li TW, Weeks KM. Structure-independent and quantitative ligation of single-stranded DNA. Anal Biochem 2006; 349(2): 242-6.
[http://dx.doi.org/10.1016/j.ab.2005.11.002] [PMID: 16325753]

[65] Orlando L. A 400,000-year-old mitochondrial genome questions phylogenetic relationships amongst archaic hominins: using the latest advances in ancient genomics, the mitochondrial genome sequence of a 400,000-year-old hominin has been deciphered. BioEssays 2014; 36(6): 598-605.
[http://dx.doi.org/10.1002/bies.201400018] [PMID: 24706482]

[66] Orlando L, Ginolhac A, Raghavan M, *et al.* True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res 2011; 21(10): 1705-19.
[http://dx.doi.org/10.1101/gr.122747.111] [PMID: 21803858]

[67] Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 2011; 27(15): 2153-5.
[http://dx.doi.org/10.1093/bioinformatics/btr347] [PMID: 21659319]

[68] Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc 2013; 8(4): 737-48.
[http://dx.doi.org/10.1038/nprot.2013.038] [PMID: 23493070]

[69] Meyer M, Fu Q, Aximu-Petri A, *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 2014; 505(7483): 403-6.
[http://dx.doi.org/10.1038/nature12788] [PMID: 24305051]

[70] Green RE, Krause J, Ptak SE, *et al.* Analysis of one million base pairs of Neanderthal DNA. Nature 2006; 444(7117): 330-6.
[http://dx.doi.org/10.1038/nature05336] [PMID: 17108958]

[71] Wall JD, Kim SK. Inconsistencies in Neanderthal genomic DNA sequences. PLoS Genet 2007; 3(10):

1862-6.
[http://dx.doi.org/10.1371/journal.pgen.0030175] [PMID: 17937503]

[72] Noonan JP, Coop G, Kudaravalli S, *et al.* Sequencing and analysis of Neanderthal genomic DNA. Science 2006; 314(5802): 1113-8.
[http://dx.doi.org/10.1126/science.1131412] [PMID: 17110569]

[73] Star B, Nederbragt AJ, Hansen MH, *et al.* Palindromic sequence artifacts generated during next generation sequencing library preparation from historic and ancient DNA. PLoS One 2014; 9(3): e89676.
[http://dx.doi.org/10.1371/journal.pone.0089676] [PMID: 24608104]

[74] Briggs AW, Stenzel U, Johnson PL, *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 2007; 104(37): 14616-21.
[http://dx.doi.org/10.1073/pnas.0704665104] [PMID: 17715061]

[75] Meyer M, Kircher M, Gansauge MT, *et al.* A high-coverage genome sequence from an archaic Denisovan individual. Science 2012; 338(6104): 222-6.
[http://dx.doi.org/10.1126/science.1224344] [PMID: 22936568]

[76] Overballe-Petersen S, Orlando L, Willerslev E. Next-generation sequencing offers new insights into DNA degradation. Trends Biotechnol 2012; 30(7): 364-8.
[http://dx.doi.org/10.1016/j.tibtech.2012.03.007] [PMID: 22516743]

[77] Green RE, Briggs AW, Krause J, *et al.* The Neandertal genome and ancient DNA authenticity. EMBO J 2009; 28(17): 2494-502.
[http://dx.doi.org/10.1038/emboj.2009.222] [PMID: 19661919]

[78] Green RE, Malaspinas AS, Krause J, *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 2008; 134(3): 416-26.
[http://dx.doi.org/10.1016/j.cell.2008.06.021] [PMID: 18692465]

[79] Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Hab Protoc 2010; 2010 (6).
[http://dx.doi.org/10.1101/pdb.prot5448]

[80] Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 2012; 52(2): 87-94.
[http://dx.doi.org/10.2144/000113809] [PMID: 22313406]

[81] Seguin-Orlando A, Schubert M, Clary J, *et al.* Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. PLoS One 2013; 8(10): e78575.
[http://dx.doi.org/10.1371/journal.pone.0078575] [PMID: 24205269]

[82] Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 2013; 29(13): 1682-4.
[http://dx.doi.org/10.1093/bioinformatics/btt193] [PMID: 23613487]

[83] Cooper A, Poinar HN. Ancient DNA: do it right or not at all. Science 2000; 289(5482): 1139.
[http://dx.doi.org/10.1126/science.289.5482.1139b] [PMID: 10970224]

[84] Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet 2001; 2(5): 353-9.

[http://dx.doi.org/10.1038/35072071] [PMID: 11331901]

[85] Stiller M, Green RE, Ronan M, *et al.* Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. Proc Natl Acad Sci USA 2006; 103(37): 13578-84.
[http://dx.doi.org/10.1073/pnas.0605327103] [PMID: 16938852]

[86] Allentoft ME, Collins M, Harker D, *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc Biol Sci 2012; 279(1748): 4724-33.
[http://dx.doi.org/10.1098/rspb.2012.1745] [PMID: 23055061]

[87] Smith CI, Chamberlain AT, Riley MS, Stringer C, Collins MJ. The thermal history of human fossils and the likelihood of successful DNA amplification. J Hum Evol 2003; 45(3): 203-17.
[http://dx.doi.org/10.1016/S0047-2484(03)00106-4] [PMID: 14580590]

[88] Geigl EM. Inadequate use of molecular hybridization to analyze DNA in Neanderthal fossils. Am J Hum Genet 2001; 68(1): 287-91.
[http://dx.doi.org/10.1086/316948] [PMID: 11115383]

[89] Barta JL, Monroe C, Teisberg JE, Winters M, Flanigan K, Kemp BM. One of the key characteristics of ancient DNA, low copy number, may be a product of its extraction. J Archaeol Sci 2014; 46: 281-9.
[http://dx.doi.org/10.1016/j.jas.2014.03.030]

[90] Skoglund P, Northoff BH, Shunkov MV, *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci USA 2014; 111(6): 2229-34.
[http://dx.doi.org/10.1073/pnas.1318934111] [PMID: 24469802]

[91] Gansauge MT, Meyer M. Selective enrichment of damaged DNA molecules for ancient genome sequencing. Genome Res 2014; 24(9): 1543-9.
[http://dx.doi.org/10.1101/gr.174201.114] [PMID: 25081630]

[92] Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. Improving access to endogenous DNA in ancient bones and teeth. Sci Rep 2015; 5: 11184.
[http://dx.doi.org/10.1038/srep11184] [PMID: 26081994]

CHAPTER 10

Complicities Between Forensic Anthropology and Forensic Genetics: New Opportunities for Genomics?

Eugénia Cunha[1,*] and **Manuela Lima**[2,3,4]

[1] Departamento de Ciências da Vida, Centro de Ecologia Funcional, Universidade de Coimbra, Coimbra, Portugal

[2] Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal

[3] Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

[4] Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal

Abstract: Human genome projects have been generating a vast amount of data which is impacting several areas, and forensic anthropology is no exception. In this chapter the interrelation between forensic anthropology and forensic genetics is highlighted, and the potential role that genomics can have in forensic science is addressed. The main sub-areas to which forensic anthropology can call the expertise of forensic genetics are listed and discussed. Genome-wide studies have recently started to be used to generate data which can efficiently aid forensic anthropologists; noteworthy, epigenetic analysis has also demonstrated its potential of application to questions that are posed to forensic anthropology. We argue that the partnership between forensic anthropology and forensic genetics is essential for stepping further in forensic sciences and that recent genomic tools have the potential to efficiently resolve questions left unanswered by genetics.

Keywords: Ancestry analysis, Forensic anthropology, Forensic genetics, Next generation sequencing, Positive identification, Sex diagnosis, Whole genome studies.

[*] **Corresponding author Eugénia Cunha:** Laboratory of Forensic Anthropology, Department of Life Sciences, Centre of Functional Ecology, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal; Tel/Fax: 351 239 240711; E-mail: cunhae@uc.pt

10.1. INTRODUCTION: MAIN ATTRIBUTIONS OF FORENSIC ANTHROPOLOGY

Nowadays the key to successful forensic expertise relies on interdisciplinary work. This means that we cannot work alone but, instead, well integrated in multidisciplinary teams. The aim of the present chapter is to highlight the interrelation between two important forensic fields, namely forensic anthropology (FA) and forensic genetics, addressing the potential role that recent genomic tools can have in forensic science. We argue that forensic anthropologists and forensic geneticists are allies and not rivals, in this respect disagreeing with Cabo (2012) [1]. In order to contextualize in what way FA needs forensic genetics, a brief introduction to the main attributions of forensic anthropology is here provided.

Undoubtedly, FA is one of the forensic disciplines which have been growing more during the last decade. This is partially due to the recognition that this expertise has been crucial to the resolution of many cases, not only routine ones, but also those arising in mass disasters and crimes against humanities scenarios. Nowadays, and in opposition to what was going on two decades ago, forensic anthropology is dealing with bodies in several states of preservation; the days when it only dealt with skeletonized remains are over. Furthermore, forensic anthropologists are also dealing with living individuals. This means that the framework of this discipline has significantly increased. The potential of human remains is now much more exploited and we can get more information even from a bone fragment. This does not mean that we are doing more with less; on the contrary, we now use more tools from several sources to decode the information saved in the human remains. And one of these tools is called genetics.

Although they are not equally important in all kinds of situations where FA can be called upon, the two main goals of a forensic anthropology expert are to identify and to assist in the determination of cause, mechanism and manner of death. All the expertise starts on the scene and without context there is no case. Therefore, fieldwork is paramount to the good resolution of a case. The basic questions arrive at the beginning: whether the remains are bones and whether the bones are human.

Within identification, there is a reconstructive phase where the victim biological profile is assessed, namely sex, stature, age at death estimation and ancestry. Afterwards, identity factors are considered: unique medical conditions and anatomical variants are searched in order to prove the uniqueness of a certain individual. Then, the comparative process of identification arrives. It is where ante and post mortem data are compared and confronted; if a match is achieved a positive identification is reached.

Cause and manner of death will mainly rely on the analysis of bone perimortem lesions. This task should be done in conjunction with the forensic pathologist, who is the only expert allowed to sign the obituary certificate. Forensic anthropologists are experts in the decoding the language of fractures and can, therefore, differentiate between ante and post trauma as well as from perimortem fractures. Moreover, they can assess the mechanism, throughout the differentiation between gunshot wounds, sharp force trauma, blunt force trauma, blast trauma or heat induced trauma. There are then secondary goals, such as the evaluation of time since death, to which FA can give paramount insights.

10.2. WHEN DOES FORENSIC ANTHROPOLOGY NEEDS GENETICS?

We can summarize the sub areas in which FA can call the expertise of forensic genetics as follows:

10.2.1. To Determine whether the Remains are Human

When dealing with small bone fragments there are situations where the macroscopic evaluation of the remains is not enough to identify the species. In these cases the interpretation is not straightforward and two alternatives can then be followed: perform histological analysis or undertake a genetic study.

When long bone epiphyses are absent or when dealing with very small skull fragments, cortical thickness and other features might not be sufficient to establish the diagnostic of Human, that is, to state that the remains match known human reference specimens to the exclusion of other reasonable possibilities [2]. In cases where the organic part is still preserved, DNA can be extracted from both bone and teeth, allowing the determination with certainty of the human origin. Polymerase-chain reaction (PCR) based techniques have been determinant to the achievement of identification on the basis of a bone fragment. Yet, there are some drawbacks, such as the invasiveness of the process, since the analyzed bone fragments will be reduced to dust. In general, whereas genetic analyses are invasive, the anthropological ones are not.

10.2.2. In Crime Scenes and Other Particular Scenarios

It is well known that all the process of identification starts in the scene where the remains are found. It is also recognized that the excavation or recovery of the remains is always an invasive process, which means that once touched, nothing can return back to exactly the original condition. Thus, in some particular situations, the recovery of DNA samples from the victims still in the scene might be very useful. In crime scenes the usefulness is evident, since DNA from the aggressor(s) can be found in the victim's body or in the context. For instances, if a gun is found nearby a dead body, DNA can be taken from the gun itself. It has to be mentioned that when bodies are in advanced state of decomposition it is hard for forensic anthropology to discuss medical legal etiology. To decide whether it was an accident, a suicide or a homicide, the context is paramount. Thus, in these cases, genetics can strengthen the context and get other sources of evidence. Other paradigmatic examples are mass disasters, such as a plane crash. In such situations, bodies or body parts are analyzed by forensic anthropologists; when recovering such bodies or body parts, DNA samples can be taken and processed immediately; this will fasten the analysis and undoubtedly contribute to more positive identifications.

One has to bear in mind that an accurate field work will help to keep chain of custody, without which a case can be canceled. In other words, a reliable recovery of the human remains by the forensic anthropologist will contribute to a reliable genetic analysis. There is a research chain and both forensic anthropology and genetics are irreplaceable links of that chain.

10.2.3. Performing Sex Diagnosis

It is known that the morphological analysis of sub-adult skeletons does not allow performing an accurate sex diagnosis. Only after puberty secondary sexual features will be visible. Therefore, for children and teenager bone remains the solution is to ask for a genetic analysis, which will routinely analyze the amelogenin gene; differences between the X and the Y chromosome versions of this gene (AMELX and AMELY, respectively) enable it to be used in sex determination of unknown human samples [3]; sex determination is based on the use of specific primers, which enable the distinction between amplicons derived from the X or the Y chromosome. The alternative of performing sex diagnosis using molecular tools is also the best practice when adult remains have a fragmentary nature.

Erroneous sex determination based on the amelogenin gene has been demonstrated as rare; for example, the study of Frances and collaborators, which has analyzed 1224 individuals, has shown that the concordance between referred sex and amelogenin test was of 99.84% [4]. Reports of anomalous amelogenin alleles include rare cases of deletions on the amelogenin gene on the Y chromosome, causing a male DNA sample to look like a female sample [5].

Sometimes contamination is hard to rule out, mainly when the samples are degraded. In such situations positive sex identifications might not be possible.

10.2.4. Performing Ancestry Analysis

Ancestry analysis is one of the areas where the cooperation between FA and genetics can be more fruitful. With globalization the evaluation of the geographic origin, the so called ancestry (Cavalli–Sforza research turned the application of race to the human species unfeasible), is becoming more complex. This happens despite the advances in the analysis of this parameter using both metric and non-metric approaches in the anthropological examination [6]. The solution relies on a mixt analysis, a double one, combining anthropology with genetics. When anthropological analysis raises a suspicion of a certain geographic area (Africa, Asia, or Europe), specific molecular markers, known as ancestry informative markers (AIMs) can have the answer. Numerous molecular analyses using combinations of single nucleotide polymorphisms (SNPs), short tandem repeats (STRs), variable number of tandem repeats (VNTRs) or even certain insertions/deletions (INDELS) indicate strong molecular patterning in worldwide samples, allowing an accurate classification of groups, despite large amounts of within region variation [7]. For that purpose, population genetics studies are essential, since we cannot find the origin of one person in a database if the respective geographic region is not represented; as stated by Callaway [8] "You can't tell someone they can trace ancestry to a certain region if that region has never been studied".

10.2.5. Identifying Specific Bacteria

Bone lesions and pathologies diagnosed on the basis of skeletal remains can work as personal identifiers. In some instances the diagnosis is not clear cut solely on the basis of the anthropological examination and in those cases the tools of forensic genetics can be a step further. Thus, in some occasions, when there is the suspicion of a victim/s having suffered from a particular disease (such as tuberculosis or leprosy) and the bone signs are equivocal, a genetic analysis can be performed. Virtually all diseases caused by agents such as bacteria, fungi and virus, can be targeted throughout DNA analysis.

The detection of some diseases by means of bacterial DNA can, moreover, also give a contribution to the knowledge of the cause of death. It is also very hard to determine the cause of death on the basis of skeletal lesions, with exception of the traumatic osseous ones, when located nearby vital organs. Yet, if the presence of a bacteria or virus can be proved, it might be possible to demonstrate that it could had been at the origin of death, even if only indirectly.

10.2.6. Performing a Positive Identification

The main advantage of forensic anthropology for victim identification is linked to its ability to shorten and simplify some steps for DNA analysis by providing a biological profile from the bones [1]. By now, DNA by itself does not provide all the information necessary for establishing positive identification.

In many instances, forensic anthropology expertise's significantly decreases the number of suspects for a possible match. This is possible because exclusions can be made on the basis of several anthropological parameters/features, such as stature and age at death. Presorting based on biological profile reduces the number of comparative samples to be obtained and analyzed. Thus, with FA, at the end, there might be two or three possible matches which will then be solved by genetics. To test a single bone sample from a skeleton and compare it to two or more family samples is one of the final steps in the identification process of decomposed human remains. In this respects it is worthwhile to cite Cabo [1]: "The extraordinary contribution of forensic anthropologists in high profile mass identification efforts such as the various human rights investigation teams across the world seem to exemplify how fruitful can be the marriage of DNA analysis and forensic anthropology. It has to be emphasized that there are not enough libraries of DNA profiles and thus, as the process is mandatorily comparative, without it, no positive identification can be achieved. By now, CODIs includes mostly offenders and arrestees" (Cabo, 2012, 454) [1].

DNA based methods used for identification in forensics are primarily based on STR loci analysis; the high level of polymorphism exhibited by this type of markers and their amenability for standardized multiplexing are amongst the arguments that justify its generalized use in identification. DNA analysis of ancestry informative markers (AIMs) and physical trait markers from biological stains can also help provide investigative leads in cases without suspects. Moreover, whilst trying to achieve a facial approximation, also called facial reconstruction, modern genetics can contribute, allowing the knowledge of details namely skin and eye color. In this respect, SNP analysis can lead to predictions about such physical traits [9, 10]. We should also bear in mind that the cases related with suspicious and doubtful identifications, as well as paternity cases involving individuals who have already died, can be solved throughout the exhumation of the concerned bodies. Again, in these cases, the forensic anthropologist selects the most appropriate bone and tooth samples (normally the femur and the molars), within the bodies in advanced state of decomposition, which will then be sent to the genetic laboratory for DNA extraction.

10.3. AN EMERGENT ROLE FOR GENOMICS IN FORENSIC SCIENCES?

Subscribing Cabana and co-authors (2013, 449) [11] "Anthropological use of molecular data has grown tremendously in the last decades". Advances in genetic technology first spurred the onset of a molecular revolution in the 1990´s, with the advent of PCR, and later, the genomic revolution of the 2000s, with the high-throughput sequencing [11]. A complete version of our genome was released in 2004 [12]. The Human Genome Project provided a reference sequence of the genome, setting the stage for several genomic projects, namely the HapMap and the 1000 Genomes Projects (for more information on Human genomic projects see Chapter 1). Simultaneously with the implementation of these projects there was a huge development of population databases; we have now an unprecedented way to advance in our knowledge by means of genomics. Citing Cabo (2012, 456) [1] from 1990 to 2008, the time required to sequence a full human genome was reduced by 7800% and the cost by 600000%. This offers a clear illustration of the speed at which DNA techniques are evolving and that we may expect more from DNA analysis in the future.

The current forensic genetics tests mainly rely on the amplification of DNA by polymerase-chain reaction (PCR), which can be routinely followed by the resolution of fragments using capillary electrophoresis with fluorescent tagging (STR analysis) or automated sequencing. Current forensic genetics tools are considered to be highly accurate and reproducible, but present, nevertheless, their own set of problems, which genomics could help to resolve. We provide some examples of questions posed to forensic anthropology and forensic genetics to which genomics can efficiently contribute.

The standard forensic genetics identification tests use multiplex PCR-based systems in which several loci are co-amplified in a single reaction, amplicons being resolved using capillary electrophoresis, as previously referred. This sort of methodology is considered of low throughput, not providing nucleotide sequence information and therefore failing to disclose eventual intra-repeat variation. Standard multiplexes have, furthermore, limitations related with available fluorochromes that limit the number of markers that can be simultaneously analyzed. In particular, mixtures comprising DNA of several individuals remain a challenge for standard methodology of forensic identification. Several studies have been showing that Next-Generation sequencing (NGS) can efficiently contribute to solve such complex situations. Bornman and collaborators [13] used massive parallel sequencing technology to the genotyping of the 13 Combined DNA Index System (CODIS) STR loci plus the amelogenin gene. The method applied showed that both individual and mixture samples were efficiently resolved. Another example of the application of high throughput methods to forensic science is the work of Zeng and co-authors [14]; using massively parallel sequencing these authors studied 23 STR loci plus amelogenin, obtaining accurate profiles.

The ability to get insights into personal characteristics as well as to produce inferences concerning ancestry has been provided by commercial panels of single nucleotide polymorphisms, commonly known as SNaPshots, which are single base extension assays. Because such assays are limited on what concerns the number of SNPs that can be simultaneously analyzed, NGS, where thousands of SNPs can be simultaneously studied, can efficiently contribute. Several studies have tested the capacity of genomic techniques for forensic SNP analysis; the need to further refine specific technical issues (such as establishing thresholds for background levels of sequence output) is, nevertheless, highlighted in various works, such as Borsting and Morling [15] and Daniel *et al.* [16].

Genome wide analysis has already been used to enlighten specific situations in which identification is needed. In a study performed to clarify the authenticity of a decorated gourd, allegedly containing the blood from the French King Louis XVI, Olalde and collaborators [17] presented evidence against this allegation; part of the evidence had a functional nature, related with both height and eye color. In fact, the complete sequencing of the genome of the sample contained in the gourd failed to find an excess of alleles associated with high stature and supported the prediction of brown eye color, whereas Louis XVI is known to have had blue eyes.

Epigenetic analysis, defined as the study of alterations in gene function caused by mechanisms other than changes in the DNA sequence [18] have recently been displaying potential for forensic identification. The hypothesis of using methylation patterns to provide information useful for sex determination, in the context of forensic genetics, had already been suggested by Naito and collaborators [19] that presented a simple procedure for sex determination based on the methylation status of the X-chromosome locus *DXZ4*. This hypothesis is now well consolidated in several global studies, such as the study from Zhu *et al.*

[20]. Epigenetic analysis has, furthermore revealed itself as potential useful in yet another aspect that is crucial for FA: age estimation. In fact, it is established that upon the recovery of a skeleton, several age-related morphological changes can be studied to provide an estimation of age at death [21]. Noteworthy, recent studies predominantly based on monozygotic twins have demonstrated that the methylation status of several genes can be of predictive value for age estimation. Globally, studies have been showing that by studying methylation in GpG islands of a selected number of loci it is possible to generate models that can predict, with high accuracy, an individual's age [22, 23]. Selection of CpG sites and adequate assay designs are considered to be critical points in the potential forensic application of methylation profile analysis, whose resolution will warrant a place for epigenetic analysis in the forensic sciences arena.

CONCLUDING REMARKS

Subscribing Dirkmaat *et al.* [24], understanding the impact of DNA analysis in forensic anthropology is critical. No doubt that the accuracy of molecular analysis changed the human identification process and that co- joint approach improved the results and impact of both disciplines under discussion.

If forensic anthropology, forensic genetics and hopefully genomics are able to develop the necessary complicities, a significant number of unsolved forensic cases can find a solution. It has to be emphasized that the sequence/order of the expertise should be followed. Thus, in a routine case, where there is no suspicion on the victim identification, the normal procedure is to start the exam by forensic anthropology and only after to proceed to genetic expertise. It is to be bear in mind that FA can also benefit forensic genetics. In fact, there are parameters such as stature and age at death to which currently genetics cannot give an accurate contribution, although there is a potential to address these issues; the present insights need further studies to be validated. So, only after having a biological profile the confrontation with the missing persons list should be done. Only after performing the exclusion of some persons from the missing persons list, and, therefore, reducing to a lower number of suspects, DNA analysis should be performed. The input that genomics can bring to FA is far from being exploited. The potential of genomics is quite huge and we are still far from depleting it; we argue that we are still giving the firsts steps, and the case of epigenetic clocks is a good example of that. To maximize the complicities, more dialogue between the involved experts is needed. As stated by Dirkmaat *et al.* [24] DNA analysis influenced the evolution of forensic anthropology as it diversified its goals and attributions and refined its theoretical framework. In all, the sciences under analysis are indeed complementary ones but more brainstorming is desirable.

CONFLICT OF INTEREST

The authors confirm that they have no conflict of interest to declare for this publication.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Cabo LL. DNA analysis and the classic goal of Forensic Anthropology. In: Dikmaat DC, Ed. A companion to forensic anthropology. 1st ed. Chichester: Blackwell Publishing Ltd 2012; pp. 447-61. [http://dx.doi.org/10.1002/9781118255377.ch22]

[2] Scientific Group of Forensic Anthropology. Available from: http://www.swganth.org/ 2013.

[3] Stone AC, Milner GR, Pääbo S, Stoneking M. Sex determination of ancient human skeletons using DNA. Am J Phys Anthropol 1996; 99(2): 231-8. [http://dx.doi.org/10.1002/(SICI)1096-8644(199602)99:2<231::AID-AJPA1>3.0.CO;2-1] [PMID: 8967324]

[4] Francès F, Portolés O, González JI, *et al.* Amelogenin test: From forensics to quality control in clinical and biochemical genomics. Clin Chim Acta 2007; 386(1-2): 53-6. [http://dx.doi.org/10.1016/j.cca.2007.07.020] [PMID: 17716640]

[5] Santos FR, Pandya A, Tyler-Smith C. Reliability of DNA-based sex tests. Nat Genet 1998; 18(2): 103. [http://dx.doi.org/10.1038/ng0298-103] [PMID: 9462733]

[6] Cunha E, Ortega PA. Cómo los antropólogos forenses evalúan la ancestría? In: Sanabria MC, Ed. Patología y Antropología Forense de la muerte – La investigación científico judicial de la muerte y la tortura, desde las fosas clandestinas hasta la audiencia pública. Bogotá, D.C., Colombia: Forensic Publisher 2016.

[7] Reno J, Marcus D, Leary ML, Samuels JE. The future of forensic DNA testing: Predictions of the research and development working group. A Report from National Commission on the future of DNA evidence Report No: NCJ 183697 Washington, DC: Department of Justice (US), Office of Justice Programs November 2000.

[8] Callaway E. Ancestry testing goes for pinpoint accuracy. Nature 2012; 486(7401): 17.

[http://dx.doi.org/10.1038/486017a] [PMID: 22678260]

[9] Bulbul O, Filogher G, Altuncul H, *et al.* A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. Forensic Sci Int Genet 2011; 3(1): e500-1.
[http://dx.doi.org/10.1016/j.fsigss.2011.10.001]

[10] Pneuman A, Budimlija ZM, Caragine T, Prinz M, Wurmbach E. Verification of eye and skin color predictors in various populations. Leg Med (Tokyo) 2012; 14(2): 78-83.
[http://dx.doi.org/10.1016/j.legalmed.2011.12.005] [PMID: 22284939]

[11] Cabana GS, Hulsey B, Pack F. Molecular methods. In: Dignagi E, Moore M, Eds. Research Methods in Human skeletal biology. Amsterdam: Elsevier Academic 2013; pp. 449-80.
[http://dx.doi.org/10.1016/B978-0-12-385189-5.00016-9]

[12] Finishing the euchromatic sequence of the human genome. Nature 2004; 431(7011): 931-45.
[http://dx.doi.org/10.1038/nature03001] [PMID: 15496913]

[13] Bornman DM, Hester ME, Schuetter JM, *et al.* Short-read, high-throughput sequencing technology for STR genotyping. Biotech Rapid Dispatches 2012; 2012: 1-6.
[http://dx.doi.org/10.2144/000113857]

[14] Zeng X, King J, Hermanson S, Patel J, Storts DR, Budowle B. An evaluation of the powerseq™ auto system: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing. Forensic Sci Int Genet 2015; 19: 172-9.
[http://dx.doi.org/10.1016/j.fsigen.2015.07.015] [PMID: 26240968]

[15] Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet 2015; 18: 78-89.
[http://dx.doi.org/10.1016/j.fsigen.2015.02.002] [PMID: 25704953]

[16] Daniel R, Santos C, Phillips C, *et al.* A SNaPshot of next generation sequencing for forensic SNP analysis. Forensic Sci Int Genet 2015; 14: 50-60.
[http://dx.doi.org/10.1016/j.fsigen.2014.08.013] [PMID: 25282603]

[17] Olalde I, Sánchez-Quinto F, Datta D, *et al.* Genomic analysis of the blood attributed to Louis XVI (17541793), king of France. Sci Rep 2014; 4: 4666.
[http://dx.doi.org/10.1038/srep04666] [PMID: 24763138]

[18] Vidaki A, Daniel B, Court DS. Forensic DNA methylation profiling potential opportunities and challenges. Forensic Sci Int Genet 2013; 7(5): 499-507.
[http://dx.doi.org/10.1016/j.fsigen.2013.05.004] [PMID: 23948320]

[19] Naito E, Dewa K, Yamanouchi H, Takagi S, Kominami R. Sex determination using the hypomethylation of a human macro-satellite DXZ4 in female cells. Nucleic Acids Res 1993; 21(10): 2533-4.
[http://dx.doi.org/10.1093/nar/21.10.2533] [PMID: 7685086]

[20] Zhu ZZ, Hou L, Bollati V, *et al.* Predictors of global methylation levels in blood DNA of healthy subjects: a combined analysis. Int J Epidemiol 2012; 41(1): 126-39.
[http://dx.doi.org/10.1093/ije/dyq154] [PMID: 20846947]

[21] Lynnerup N, Kjeldsen H, Zweihoff R, Heegaard S, Jacobsen C, Heinemeier J. Ascertaining year of birth/age at death in forensic cases: A review of conventional methods and methods allowing for

absolute chronology. Forensic Sci Int 2010; 201(1-3): 74-8.
[http://dx.doi.org/10.1016/j.forsciint.2010.03.026] [PMID: 20399051]

[22] Bocklandt S, Lin W, Sehl ME, *et al.* Epigenetic predictor of age. PLoS One 2011; 6(6): e14821.
[http://dx.doi.org/10.1371/journal.pone.0014821] [PMID: 21731603]

[23] Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. Aging (Albany, NY) 2011; 3(10): 1018-27.
[http://dx.doi.org/10.18632/aging.100395] [PMID: 22067257]

[24] Dirkmaat DC, Cabo LL, Ousley SD, Symes SA. New perspectives in forensic anthropology. Am J Phys Anthropol 2008; 137 (Suppl. 47): 33-52.
[http://dx.doi.org/10.1002/ajpa.20948] [PMID: 19003882]

SUBJECT INDEX

Manuela Lima, Amanda Ramos & Cristina Santos (Eds.)