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

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
	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]		
	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]		
	2005	To answer fundamental questions about the origin of human populations and their migration paths.
	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]		
	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]		
	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]		

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

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

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

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

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].

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

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

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

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

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

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

SUBJECT INDEX

A

B

C

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V

X

Y