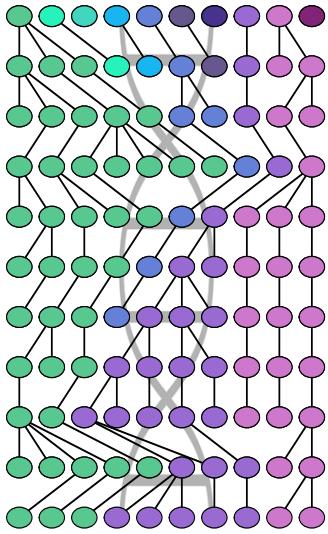
UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Unveiling the genomic history of human populations in macro- and microgeographic contexts: from Eurasia to the Americas



Nicola Rambaldi Migliore

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIV – A.A. 2018-2021

On the cover:

Scheme of a Wright-Fisher population of alleles experiencing genetic drift produced using the R function *coalescent.plot()* (https://github.com/liamrevell/learnPopGen/blob/master/R/evolution.R) from Revel, L.J. "*learnPopGen*: An R package for population genetic simulation and numerical analysis", *Ecol Evol*. 2019 Jul; 9(14): 7896–7902.

UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Unveiling the genomic history of human populations in macro- and microgeographic contexts: from Eurasia to the Americas

Nicola Rambaldi Migliore

Supervised by Prof. Alessandro Achilli

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXIV – A.A. 2018-2021

Table of Contents

1. *Abstract*

The genomic histories of human populations have been recently enriched by modern and ancient DNA data. The new field of archaeogenomics, which aims at a parallel analysis of these datasets, is providing additional details on the complexity of the past of humankind with multiple migrations, demographic spreads, and admixture events worldwide. As for the molecular targets, initial studies were based on the analysis of uniparental systems, the mitochondrial DNA (mtDNA) and the Y chromosome. Nowadays, genetic surveys have extended to entire genomes from both modern and ancient individuals.

The projects presented in this thesis combine one or several of these features and investigate the genetic histories of human populations from Eurasia and the Americas in different geographic contexts.

In the first project [1], based on a microgeographic setting, the diachronic comparison of 198 complete mtDNAs (selected based on the variation of 545 modern mtDNA control regions) with 19 ancient mitogenomes from the Umbria region in Italy highlighted a long and complex history of migrations from different sources and in different times, with genetic continuity between ancient and modern individuals in the eastern part of the region.

Moving to eastern Asia, the second work [2] investigated the variation of 2,420 modern mtDNAs (147 of which at the level of complete mitogenomes) from all over Mongolia. The results showed a clear genetic differentiation in the region. East Asian contributions suggested continuous connections with neighboring populations until recent times. Western Eurasian mitogenomes highlighted two major demographic changes, the first of which occurring during the Late Pleistocene and suggesting a connection with post-glacial repopulation events from western Eurasian refuges. Lastly, a comparison with Mongolian ancient individuals suggested a link between the occurrence of some mtDNA lineages and the timeframe and geographic path of the Silk Route.

After traversing the strait of Bering towards the Americas, in the third work [3], also set on a macrogeographic context, we built a country-wide dataset of 2,021 mtDNAs from the present-day general population of Mexico. The findings confirmed that the genetic impact of European conquest was small in terms of maternal lineage introgression, with preservation of Indigenous mtDNA lineages, heterogeneously distributed within the country. The proportion of west Eurasian mtDNA haplogroups was found to be low, but with some exceptions, mainly restricted to the northeast. The results also pointed to a possible sex-differentiated mobility and mixture that impacted cultural as well as biological survival in Mexico. Furthermore, we also

highlighted the importance of country-wide and regional databases for forensic genetic investigations.

Moving south across the double continent, I have focused my main projects on the Isthmus of Panama, an obligatory passage for the first peopling of the Americas and an important crossroads during both the European colonization and the African slave trade. The existence of sex biases in the convergence of diverse Indigenous groups was often documented by the differential inheritance of uniparental lineages. In this paper [4], I have studied and compared the mtDNA and Y-chromosome variation of 431 individuals (301 males and 130 females) belonging to either the general population, mixed groups, or one of five Indigenous populations currently living in Panama. We found different proportions of paternal and maternal lineages in the Indigenous groups testifying to pre-contact demographic events and genetic inputs (some dated to Pleistocene times) that created genetic differentiation. We also showed that the local mtDNA gene pool (especially among the Indigenous populations) was marginally involved in post-contact admixtures, whereas the Indigenous Y chromosomes were differentially replaced, mostly by west Eurasian lineages. Moreover, we provide new estimates of the sub-Saharan African contribution to a more accurately defined general population of Panama, somehow reconciling genetic data with historical records.

In addition to the projects on mtDNA sequence data, I also had the opportunity to start working with both ancient and modern autosomal DNA data. The archaeogenomic study of Panama [5] analyzed 84 genome-wide profiles from present-day Indigenous and admixed groups, together with the first reliable 12 low-coverage ancient genomes from this region. An initial evaluation of the ancient data made it possible to address long-standing anthropological and archaeological questions regarding the possible genetic relationships among individuals buried together, testifying to the multidisciplinary feature of this work. Moreover, we identified a remarkable genomic structure within Panama. Nevertheless, Panamanian groups also showed relatedness, especially in western regions, mirroring the pre-contact cultural area of Greater Chiriquí. Fewer genetic similarities were instead identified between the Indigenous populations located in eastern Panama, known in pre-contact times as Greater Darién. In the wider continental genomic landscape, we described the presence of a specific axis of Indigenous genetic variation in the Americas, which is typical of the Isthmo-Colombian area. This component was present not only among pre-Hispanic Isthmian individuals, but also strongly characterized present-day Panamanian groups, surviving both pre-colonial demographic fluctuations and the genetic bottleneck (and admixture) caused by colonialism. Eventually, such genomic distinctiveness has been explained by considering an additional Pleistocene ancestry that further enriches the Indigenous American genetic history.

Crossing the Isthmus to South America, I was involved in a still ongoing project on the Ashaninka Indigenous population from Peru. The preliminary analyses of genome-wide data from 44 individuals highlighted an outlier behavior of the Ashaninkas within the South American genomic landscape, the occurrence of two main genetic clusters within this population, and peculiar connections with ancient Caribbean groups, a link that will need further investigation.

This thesis highlights some important aspects of human population genetics and provides a relevant contribution to the field. First, mtDNA studies are confirmed as fundamental to reconstruct phylogenies and for evolutionary and forensic genetic applications. However, uniparental systems are based on single non-recombining *loci*, representing only two ancestral paths out of the thousands that contribute to a genome. This complexity can be tackled by accompanying autosomal data to uniparental markers and by further incorporating evidence from ancient DNA, which has revolutionized the field of population genomics in the last decade, becoming an important tool to investigate the genetic histories of populations.

2. Acknowledgements

First, I would like to thank my supervisor Prof. Alessandro Achilli. He taught me a lot and has always been available to help and open to scientific discussions. Thanks for letting me work in this fascinating field of science.

Then, I would like to thank all the people in the human and animal population genomics lab. Thanks to Profs. Antonio Torroni, Ornella Semino, Anna Olivieri, Luca Ferretti, and to Dr. Viola Grugni. Thanks to Johnny, Giulia, all the students in the group and the newcomers.

A thanks goes also to all the member of the Achilli lab: Ana, Nataliia, Vittoria, Giulia, Aleksandro, and the new students.

Thanks to all the collaborators around the world that made all projects in which I have been involved possible.

A special thanks to Drs Hansi Weissensteiner and Licia Colli for having accepted to be the external reviewers of this thesis, and for their insightful comments and suggestions.

To end with the professional thanks, I would like to express my gratitude to all the people that accepted to donate their DNA to be analyzed.

Now, I would like to give heartfelt thanks to Marco, Linda, and Ale. They are the best friends and colleagues one could ask for, both back when we were all in Pavia, and now that we are spread between Italy and Ireland. I can't wait to get our next "birretta" all together!

I also want to thank my best friends in Cremona. I've known them for a lifetime and shared a lot together. It's also thanks to them I could achieve this, and many others, result (and thanks to our Gloomhaven games, that allowed me to relax, have fun, and recharge in these last tough few months).

A heartfelt thanks to my parents and my sister for having always supported me through my decisions and all the difficulties. Thanks for having allowed me to become the person I am now. None of this would have been possible without them being always there for me. A special thanks also to Lizy.

Finally, I want to thank Sara (with Frolla and Oreo). I don't even know where to start. Thanks for always being there for me, for having taught me that I am worth it, that life is beautiful despite it all. Thank you for your smile. Thanks for, as you use to say, having completed this PhD journey with me, as if you did one yourself, despite your aches and pains. Thank you for everything, even if you always scold me.

3. Abbreviations

1KGP: 1000 Genomes Project

aDNA: Ancient DNA

AMH: Anatomically modern humans

ANA: Ancestral Native Americans

ANGSD: Analysis of Next Generation Sequencing data

ANS: Ancient North Siberians

BAM: Binary Alignment Map (compressed raw mapping file)

BCE: Before common era

BCL: Binary Base Call format (raw data from Illumina sequencers)

BEAST: Bayesian Evolutionary Analysis by Sampling Trees

bp(s): Base pair(s)

BSP: Bayesian Skyline Plot

BWA: Burrows-Wheeler Aligner

CA: Correspondence Analysis

CE: Common Era

CRS: Cambridge Reference Sequence

DNA: Deoxyribonucleic acid

GATK: The Genome Analysis Toolkit

Gb: Giga bases

Hd: Haplotype diversity

HGDP: Human Genome Diversity Project

Hg(s): Haplogroup(s)

HTS: High-Throughput Sequencing

HVS: Hypervariable Segments

IAm: Indigenous American

IBD: Identity -By Descent

k: Average number of nucleotide differences

Kb: Kilo base pairs

ky: Thousand years

kya: Thousand years ago

LD: Linkage Disequilibrium

LGM: Last Glacial Maximum

LR: Long Range

MAFFT: Multiple Alignment using Fast Fourier Transform

MCMC: Markov chain Monte-Carlo

MDS: Multidimensional Scaling

ML: Maximum Likelihood

MP: Maximum Parsimony

MRCA: Most Recent Common Ancestor

MSY: Male-specific region of chromosome Y

mtDNA: Mitochondrial DNA

N_e: Effective population size

NGS: Next Generation Sequencing

NJ: Neighbor Joining

NNA: Northern Native Americans

nps: Nucleotide positions

NRY: Non-Recombining region of Y chromosome

PAR: Pseudo Autosomal Region

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PCs: Principal Components

Pi: Nucleotide diversity

rCRS: Revised Cambridge Reference Sequence

RFLPs: Restriction Fragment Length Polymorphisms

rRNA: Ribosomal RNA

SAf: Sub-Saharan African

SAM: Sequence Alignment Map (raw mapping file)

SFS: Site Frequency Spectrum

SGDP: Simon Genome Diversity Project

SMC: Sequentially Markov Coalescent

SNA: Southern Native Americans

SNP: Single Nucleotide Polymorphism

STRs: Short Tandem Repeats

tRNA: Transfer RNA

UPGMA: Unweighted Pair Group Method with Arithmetic Mean

UPopI: Unsampled Population of the Isthmus

VCF: Variant Call Format

WEu: Western Eurasian

WGS: Whole-Genome Sequencing

4. Introduction

4.1 Background on population genetics

Genomes are made of Deoxyribonucleic Acid (DNA), that contains the information needed for the life of an organism and which is passed down generation after generation.

The ultimate source of all evolutionary changes in DNA is mutations. Mutations can be of different types, mainly occurring during DNA replication before cell division, and they cause different changes in the genome. The most abundant kind of mutation is the conversion of one nucleotide into another, which in a population, when the frequency of the new nucleotide is higher than 1%, becomes a single-nucleotide polymorphism (SNP). Indels, which are insertions or deletions of short sequence fragments, are the second most abundant type of polymorphism in human genomes [6]. They are especially frequent in the context of repeated sequences of few base pairs (bps) as short tandem repeats (STRs) [7], due to polymerase slippage during replication. In addition to these small-scale changes, genomes undergo larger-scale structural variations, in which large chromosome tracts are duplicated, deleted, inverted, or translocated to a different position of the genome. Humans are diploid organisms, therefore, for each given polymorphic *locus* on the autosomes and the recombining part of the sex chromosomes, they have two alleles (variant forms of DNA sequences occurring at the same *locus*), with one allele inherited from each parent. Each pair of alleles represents the genotype at a specific genetic *locus*.

Another source of change in the genome is the process of meiosis in sexually reproducing organisms. Consequences of meiosis are segregation, independent assortment, and crossing-over. The latter, also known as meiotic recombination, results in the shuffling of genetic material between homologous chromosomes. Without recombination, DNA sequences are passed on over generations and accumulate mutations sequentially over time, as occurs for the mitochondrial genome (mtDNA) and the non-recombining region of the Y chromosome. The autosomes passed down to the offspring, after meiosis, can be a recombined version of the two chromosomes carried by the parent, and the probability by which recombination occurs in any part of the genome can vary. The closer two variants are on a chromosome, the less likely recombination will occur between them. The resulting co-segregation of these variants at a frequency greater than expected by chance is known as linkage disequilibrium (LD). The specific combination of variants that are co-inherited on a chromosome, or on a fragment of a chromosome, are called haplotypes.

These changes are the main causes of variation in the genome. DNA is thus shaped over time by evolutionary processes, which are mainly driven by two forces [8]: natural selection, the result of variation in the probabilities of transmission among individual genotypes, and genetic drift, the random variation of allele frequencies in a population [9, 10]. This random fluctuation depends on the probability of individuals carrying a specific allele at a given *locus* of leaving more or less offspring in the next generation. In a sexual population, drift also depends on only one of the two alleles at a *locus* in an individual being transmitted to the offspring. Genetic drift mainly impacts neutral alleles. A neutral polymorphism occurs when the segregating alleles at a polymorphic site have no differences in their effect on fitness. The role of drift in the evolutionary processes has been debated since the proposition of the neutral theory of molecular evolution [11, 12]. This theory is based on the assumption that most polymorphisms are neutral alleles, subject only to genetic drift and mutation. Genetic drift will, in absence of mutations, remove neutral genetic diversity, as alleles will eventually either be lost or fixed over time (Figure 1).

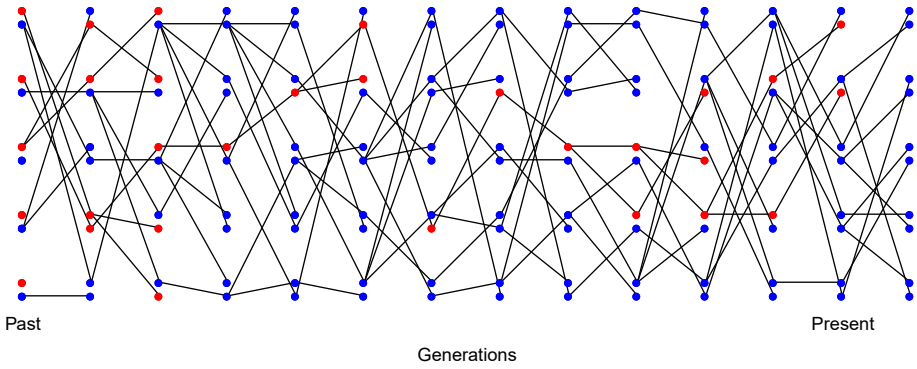


Figure 1. Loss and fixation of the red and blue alleles, respectively, over the generations in the absence of mutation. Figure adapted from (https://github.com/cooplab/popgen-notes), under the terms of the Creative Commons Attribution 3.0 Unported License (http://creativecommons.org/licenses/by/3.0/).

Another important concept in population genetics is the coalescent theory [13], which states that all alleles at a given *locus* derive from a common ancestral molecule, also referred to as the Most Recent Common Ancestor (MRCA). From this statement it follows that one can calculate the coalescence time, that is when the ancestral allele was the only one present in past generations. The coalescence time considers the probability that two alleles coalesce in the previous generation as $1/(2N_e)$. N_e is the effective population size, that is the size of an idealized constant population which matches the rate of genetic drift, or the size of the subset of the population that is actually contributing genetic material to the next generation [14]. When the N_e is small, random changes in allele frequency have a greater impact

and genetic drift occurs more rapidly. The coalescent theory has many applications in population genetics and can allow us to investigate and date events, as migrations and time to the MRCA, extrapolating this information from modern and ancient sequence data.

In this context, another useful concept is the mutation rate, which gives the probability of the occurrence of a mutation in a genome in each generation. These rates can differ among species, within the same species, and even within the same genome. Mutation rate is usually estimated as the probability of a substitution per site per generation. In humans, the mutation rate is approximately 1.25×10^{-8} per base pair (bp) per generation, although there is still a debate about this number, considering how much it can vary between different regions of the genome [15, 16]. Since DNA sequences evolve at an approximately constant rate, the number of differences between genomes is proportional to the time since they last shared a common ancestor [17]. This is the principle at the basis of the concept of the molecular clock, which can be very useful for estimating evolutionary timescales. It usually requires a calibration given by the comparison with an outgroup whose divergence time is already known from other sources.

4.2 Technologies for obtaining genomic data

At the beginnings of population genetics, the greatest advancement was the introduction of molecular tools to study variation within and between species [18]. The first molecular markers used to this aim were proteins and blood group antigen proteins, recognized with antibodies, have been used to carry out the very first studies on molecular variation [19]. Attempts to quantify genetic variability using polypeptides continued in the 1960s [20, 21] although with technical limitations, since the methodologies were only based on the assessment of protein mobility in gel electrophoresis, introduced at the end of the 1940s [22].

After the discoveries of DNA being the genetic material of the cell [23] and of the DNA structure [24, 25], it became clear that further advances would have required studies of DNA sequence variation. A step forward was the detection of DNA sequence variants by identifying mutations in sites cut by restriction enzymes, discovered in bacteria [26]. A great breakthrough occurred later on with the development of the polymerase chain reaction (PCR) [27], which enabled researchers to obtain high copy numbers of specific DNA sequences. The combination of PCR with gel electrophoresis allowed to resolve the occurrence of amplified fragments of varying length at highly variable microsatellite *loci* (or STRs). Direct sequencing of DNA started in the 1970s and the method that became predominant in the following decades was the Sanger sequencing technology [28] (Figure 2).

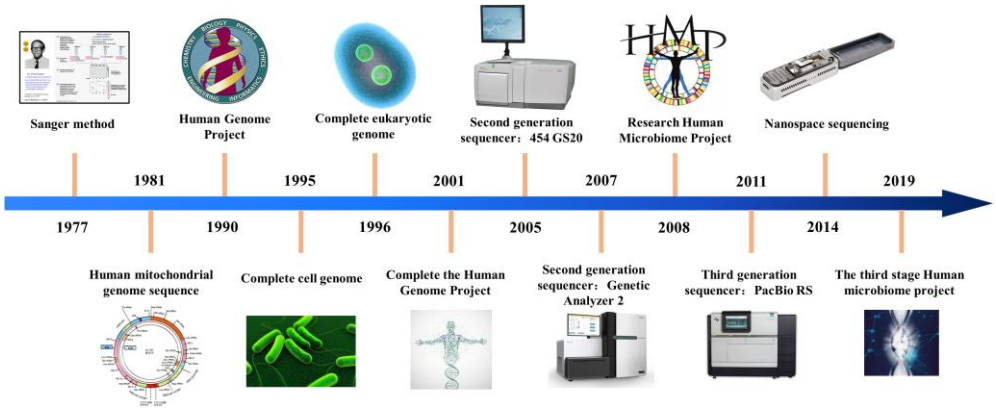


Figure 2. History of DNA sequencing technologies. Unchanged from Figure 1 in Yang et al [29], under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The Sanger method was used, for example, to sequence the first human genome in the frame of the Human Genome Project, with first incomplete assemblies being published in 2001 [30, 31]. Another technological development for the study of genetic variation was the setup of high-density microarrays. These arrays can hold large numbers of short oligonucleotide probes that target specific regions of the genome, corresponding to variants known to segregate in a population. Therefore, the genotypes at millions of polymorphic *loci* can be determined with high accuracy and reproducibility [32]. Since not all the arrays have been designed for population genetic studies, they have some drawbacks that must be considered [33]. One of these is the so-called ascertainment bias: the choice of the SNPs to include in an array is often biased towards the population(s) in which these variants were scored. Since most variants for humans have been ascertained in populations of European ancestry, when more distant populations are studied, part of the genetic variation may be lost, due to the occurrence of many monomorphic markers. To circumvent this drawback and return a less biased estimate of population variability, the Axiom™ Genome-Wide Human Origins 1 Array has been designed [34], to include ~630,000 SNPs extracted from 13 population-specific panels, analyzed in the Human Genome Diversity Project (HGDP) as well as the Neanderthal Genome Project [35]. This array was used in our genome-wide studies on the Panamanian and Ashaninka modern populations (chapters 8.1 and 8.2).

Another major breakthrough occurred with the development of the next-generation sequencing (NGS) technologies in the first decades of the 21st century [36]. NGS opened a new era for genomics, reducing both the cost and the time needed for whole-genome sequencing (WGS) at increasing coverage. This led to projects and consortia aiming at sequencing as many genomes as possible (not only humans) worldwide, such as the 1000 Genomes Project (1KGP) [37], the Simon Genome Diversity Project (SGDP)

[38], and the Human Genome Diversity Project (HGDP) [39]. In addition, large-scale databases, containing tens to hundreds of thousands of whole genomes from specific countries, are now becoming common [40–46]. The technological advancement continued beyond NGS, leading to what is known as third-generation sequencing [47]. Thanks to the long-read technology and other features, this novel technique is providing a way to study genomes at an unprecedented resolution. This led to the improvement of the human reference genome by filling many gaps [48], and to a deeper investigation of epigenetic patterns [49] and segmental duplications [50]. Lastly, NGS has also allowed a revolution in the field of ancient DNA (i.e., DNA extracted from archaeological remains), making it possible to obtain whole ancient genomes starting from 2010 [51] (see chapter 4.6).

4.3 Methods for analyzing genomic data

There is a variety of computational methods for learning about population histories from genetic data. They differ on the basis of data features, amount and type. A first and commonly used example is the phylogenetic tree, a diagram representation of the variation of groups of DNA sequences (*taxa*) deriving from a common ancestor. It is composed of branches, showing the genetic differentiation among the data, and internal nodes, that represent the MRCA of each group (Figure 3).

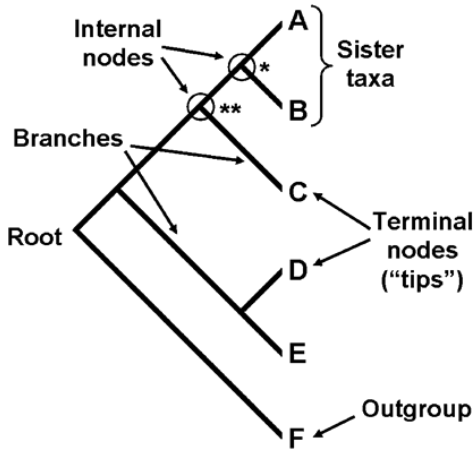


Figure 3. A schematic example of a phylogenetic tree. The terminal nodes, or tips, at the right represent the data (e.g., the species under analysis). These are connected by branches, joining at the internal nodes, which represent the MRCA of each group. For example, the internal node with one asterisk is the MRCA of species A and B, which are sister groups, since they share a MRCA not shared by the other species in the phylogeny. The outgroup is the most distantly related species in the phylogeny. Overall, the topology of the tree indicates evolutionary relatedness. Figure from Gregory [52], under the terms of the Creative Commons Attribution Noncommercial License (https://creativecommons.org/licenses/by-nc/2.0).

Phylogenetic trees can be built using different techniques. Distance-based methods make use of a pairwise distance matrix. Among these are the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and the Neighbor Joining (NJ). The UPGMA builds a unique phylogenetic tree by progressively combining the two taxa with the smallest genetic distance, while the NJ first builds a minimum evolution tree (with the smallest sum of branch lengths), and then reaches the final phylogeny by progressively establishing branches based on all the posterior distance-based iterations. Character-based methods, instead, start by assuming an explicit evolutionary model. Among them there are the Maximum Parsimony (MP) method, which generates a tree considering the smallest number of evolutionary changes, and the Maximum Likelihood (ML), that evaluates different topologies and then finds the tree explaining the data with the highest likelihood, assuming a best fitting evolutionary model [53], which can be identified using tools, such as jModelTest [54]. There are different methods which can be used to date the internal nodes of a phylogenetic tree. A very simple one, commonly used [55] for mtDNA phylogenies, assumes that mutation rate is constant across the branches, and calculates time as proportional to the average number of substitutions between a set of sequences and their MRCA. The age of the nodes can also be calculated using an ML approach, which calculates the probability of the data given the hypothesis (the tree). Another approach is the Bayesian method, determining the probability of the hypothesis (the tree) given the data. This is based on Markov chain Monte-Carlo (MCMC) simulations with prior estimates to generate the phylogeny with the maximum posterior probability. These methods are well suited to study the variability of non-recombining *loci*, such as mtDNA and Y chromosome, but phylogenetic trees are not the best way to investigate autosomal data, given that each *locus* has its own MRCA. Therefore, there are other approaches that can be used to study genome-wide autosomal data.

Several of these methods are based on allele frequencies at biallelic polymorphic sites (SNPs), assuming that populations or individuals that share ancestry will have similar allele frequencies. Principal component analysis (PCA) is a multivariate statistical method commonly used to reduce the high dimensionality of the data (in the form of a covariance matrix), by applying eigenvector decomposition and reduce the variability to principal components (PCs), that explain a large portion of the original variation [56]. It allows us to detect genetic differentiation between individuals/populations, usually by looking at the first two (most informative) PCs (Figure 4).

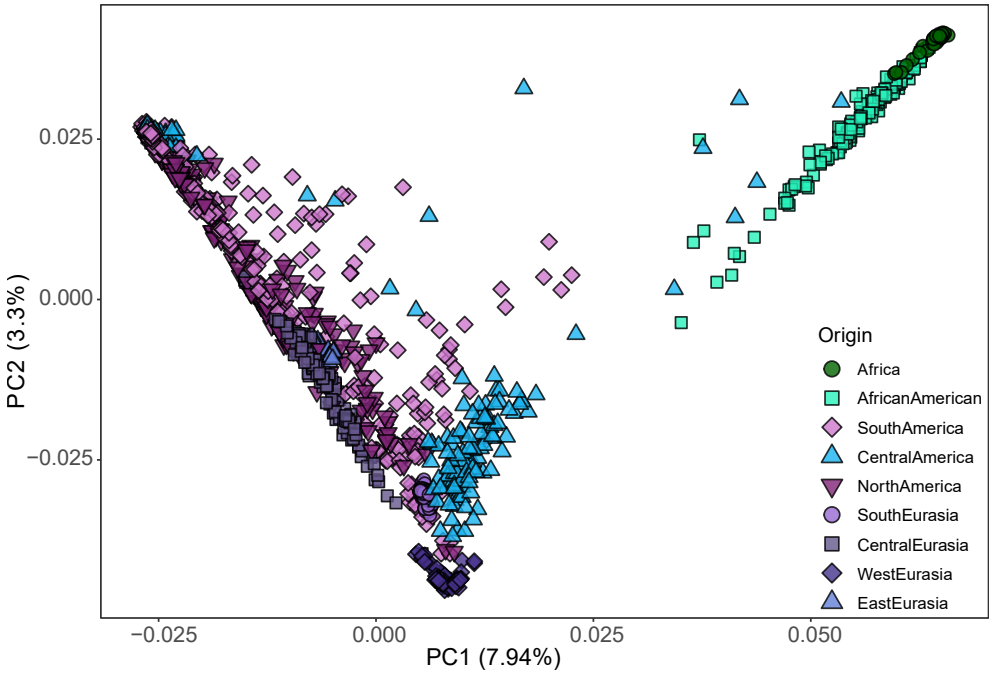


Figure 4. Example of a PCA, computed on 1,558 individuals (and 406,920 SNPs) from American and other worldwide modern populations [5].

Model-based clustering methods implemented by the STRUCTURE [57] and ADMIXTURE [58] software are also commonly used. These are based on clustering algorithms that group the data into K ancestral populations (Figure 5), or genomic components, based on genetic affinities or differences between genotypes. Each K represents a hypothetical population of individuals sharing a typical genetic pattern. These methods use unlinked or independent *loci,* generally requiring the *a priori* definition of the number of genomic components (K). The software subsequently assigns a fraction of every genome to each of the K clusters. However, choosing the optimal number of K is often challenging. Therefore, most of the software also implement tools that help selecting the best K, as the calculation of a cross-validation error implemented in the ADMIXTURE package.

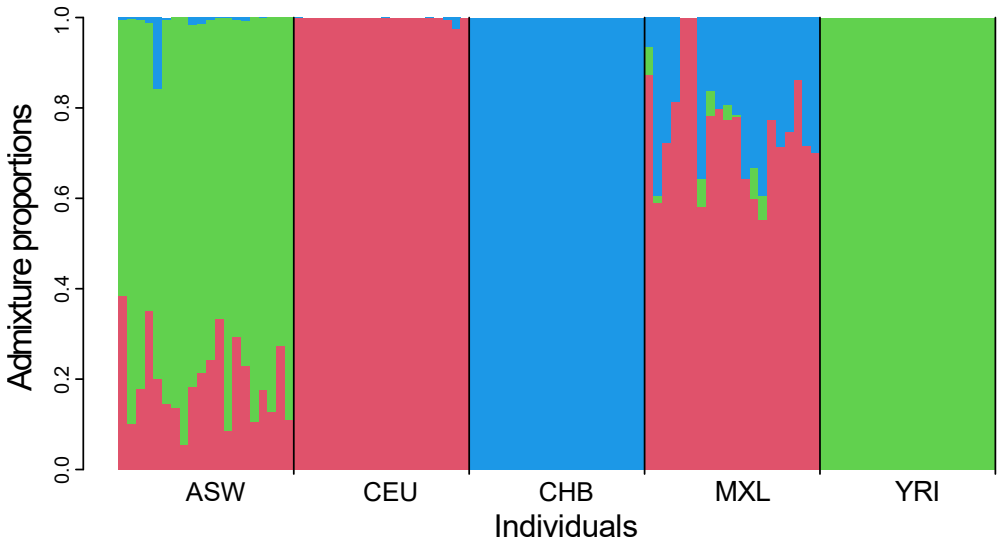


Figure 5. Example of ADMIXTURE analysis on 100 modern individuals (50,000 SNPs) from the 1000 Genomes Project [37]. ASW: US individuals with African ancestry; CEU: US individuals with European ancestry; CHB: Han Chinese; MXL: Mexican individuals; YRI: Yoruba individuals from Nigeria.

Another important value that gives information on the distance between populations is the fixation index F_{ST} [10]. This is a measure of the allele frequency differences among populations (pairwise F_{ST}). Its values range from 0 to 1, with 0 indicating populations that are freely interbreeding and 1 implying they do not share any alleles with one another, i.e., they are completely isolated and do not interbreed.

Starting from Wright's fixation index, f_2-, f_3-, and f_4-statistics [59, 60] have been developed, which investigate admixture, population structure, and genetic distance [61]. They are based on the estimation of shared genetic drift between populations, assuming that its occurrence indicates a shared evolutionary history. *F*-statistics measure allele frequency correlations to test if allele frequencies are consistent with a simple tree topology or if there is admixture in the history of populations. Starting from a simple tree topology, each branch length corresponds to the shared genetic drift among genomes within the branch. If there is admixture, the alternative model is an admixture graph which extends the phylogeny by allowing edges that represent admixture events.

The purpose of f_2-statistic is to measure how much genetic drift occurred between two populations, i.e., to measure genetic dissimilarity, such as the F_{ST}. Drift is quantified as the variance in allele frequency between the two populations, taking into account all the polymorphic *loci* across the genome that have been genotyped [59, 60]. The approach is to interpret $f_2(P_1, P_2)$ as a measure of dissimilarity between populations P_1 and P_2 (Figure 6), as a

large f_2 value implies that populations are highly diverging. Thus, the strategy is to calculate all pairwise f_2 indexes between populations, combine them into a dissimilarity matrix, and test whether that matrix is consistent with a tree. For a diversity matrix to be consistent with a tree, the length of all branches must be positive, since negative genetic drift is biologically nonsense, and thus negative branches are a violation of treeness [61]. F_3- and f_4-statistics derive from f_2-statistic. These are used to test for admixture events evaluating treeness hypothesis between three and four populations, respectively (Figure 6).

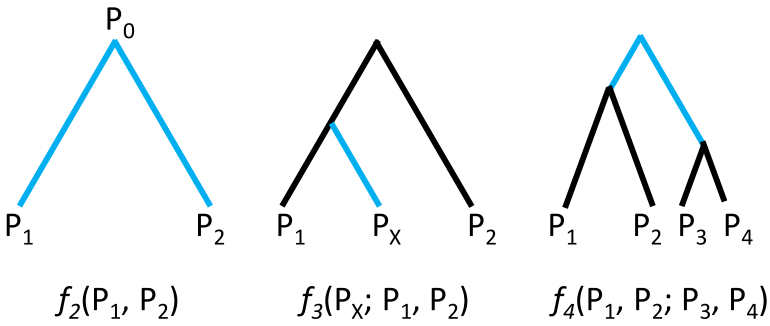


Figure 6. Schematic representations of f_2-, f_3-, and f_4-statistics.

F_3-statistics has two main applications. The first one is to identify admixture events. Considering the populations $f_3(P_x; P_1, P_2)$ (Figure 6) and p_x, p_1, p_2, as their allele frequencies, if the product $(p_x - p_1)$ $(p_x – p_2)$ is negative, then P_x descends from admixture between P_1 and P_2. Another application of f_3 is known as outgroup f_3-statistic, defined as $f_3(P_O; P_1, P_2)$ where P_O is an outgroup to the other two populations. This measures the shared genetic drift between P_1 and P_2, with respect to P_O, and high values mean long shared branches and therefore high shared drift. Thus, the higher the f_3 values, the more closely related P_1 and P_2 are.

The f_4-statistic is defined as $f_4(P_1, P_2; P_3; P_4)$ (Figure 6), and p_1, p_2, p_3 and p_4 are the respective allele frequencies. By calculating $(p_1 - p_2)$ $(p_3 – p_4)$, it is possible to test if P_1/P_2 and P_3/P_4 form clades. If the value is close to zero, the proposed tree topology is confirmed, instead if it is negative P_1 is more closely related to P_3 and P_2 to P_4. More complex scenarios can be assessed taking together all the results of all the *f*-statistics in a group of populations. This approach is implemented in software packages as qpWave, qpAdm, and qpGraph, which are used to define the number of source ancestries for a population (given a set of outgroups), the proportions of those ancestries in a population, and a model of evolutionary relationships in a set of populations, respectively [59, 60].

Other methods additionally make use of haplotypes and are called haplotype-based methods. A requirement is that the data must be phased, which implies that haplotypes must be estimated from genotype data, assigning alleles to the paternal or maternal chromosomes. The most used approach for haplotype phasing is to make use of a large reference panel of haplotypes, to which input genotypes are compared and the most likely phase is inferred [62]. To gain linkage information, haplotype-based methods also need a high density of variants, such as that offered by whole-genome data or high-density genotyping arrays.

Among these methods, the CHROMOPAINTER and fineSTRUCTURE software packages employ a framework where haplotypes are compared between individuals to assess similarity in a high-resolution fashion [63]. CHROMOPAINTER is based on the knowledge that haplotype segments shared among individuals become shorter over time due to recombination. Therefore, it allows the reconstruction of the chromosomes of a set of individuals as a series of genomic chunks inherited from another set of individuals. FineSTRUCTURE uses a MCMC clustering model based on haplotype similarity patterns. It takes CHROMOPAINTER output and identifies clusters with indistinguishable haplotype similarity profiles. Other methods aim to infer the local ancestry of haplotypes in admixed populations, by using reference panels related to the sources of admixture, such as RFmix [64], that matches segments in admixed individuals to those from the source populations. Other methods have also been developed, that infer local ancestry in a more sophisticated way [65, 66] and that can be used to date admixture events using LD patterns [65].

Populations' sizes may change over time for a variety of reasons, such as population growth, migrations, bottlenecks, and many others. Demographic studies are therefore crucial in evolutionary and population genetics. Methods used to infer demographic histories and fluctuations in the effective population size mostly require haplotype phased data. Therefore, they work best with whole-genome sequences, possibly with high coverage. Some of these methods are based on the Sequential Markovian Coalescent (SMC), such as MSMC [67] and allow reconstructing the history of a population using very few genomes or even a single one (and its two haplotypes). Other tools are based on the Site Frequency Spectrum (SFS) and therefore consider the frequency of rare and common variants to reconstruct population histories, as Momi2 [68].

4.4 Uniparental systems

Uniparentally-inherited molecular systems have been, and still are, a major tool in population genetics, even though autosomes are now important high-resolution markers to study population genetic structure and human demographic histories (Figure 7). Uniparental systems have many applications also in other fields, such as evolution, medical and forensic genetics, and genetic genealogy [69]. The maternally inherited mtDNA and the paternally inherited male specific region of the Y chromosome (MSY) are extensively used for reconstructing molecular genealogies and phylogenies [70–72]. The mode of inheritance means that they contain records of male- and female-specific population processes.

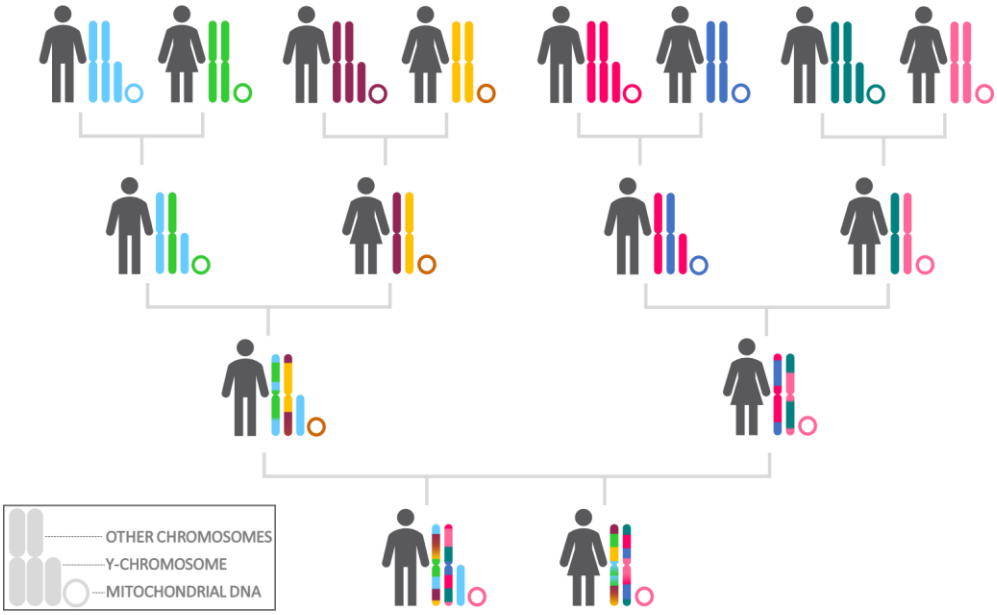


Figure 7. Overview of the human uniparental and autosomal genetic systems and their transmission over generations. Image credit to Dr. Marco R. Capodiferro.

Uniparental systems do not undergo meiotic recombination, and thus accumulate sequence variation through random mutation over time. This process of molecular divergence has given rise to the monophyletic units that are called haplogroups (Hgs), lineages that are characterized by the same combination of mutations [73]. All contemporary mtDNAs and MSY regions coalesce back to one ancestral molecule, which was present in a certain moment of the past, from which different haplogroups and sub-haplogroups split, by accumulating a series of mutations that occurred independently and accumulated differently. Since the sequential accumulation of the new mutations is relatively fast and occurred mainly while modern humans were colonizing different regions and continents, haplogroups and sub-

haplogroups tend to be restricted to specific geographic areas and populations [74]. Phylogeography [75], together with phylogenetic trees, has been extensively used to reconstruct the origin and the migrations of populations. In addition, being the coalescent approach well suited for uniparental systems, it also has been applied to date migration events at a population level and to analyze demographic changes through time.

In this PhD thesis, I will describe works on human mtDNA uniparental system. Therefore, in the following sub-chapters, I will deeper describe the mtDNA features.

4.4.1 Y chromosome

The Y chromosome is ~60 Mb in length, it plays a crucial role in sex determination at the embryo level and in the male fertility. It is transmitted only along the paternal line and does not undergo recombination [73]. In particular, the non-recombining part is called Male Specific Y region (MSY), also referred to as the Non-Recombining region of Y chromosome (NRY). Three Pseudo Autosomal Regions (PAR) are the only portions that undergo recombination with the X chromosome [76]. The MSY has a higher average SNP mutation rate than the rest of the nuclear genome. The MSY is a powerful tool in evolutionary and population genetics. Y-chromosome polymorphic variants are biallelic (SNPs) and multiallelic (STRs). Both SNPs and STRs have been and still are widely used in evolutionary and population genetic studies. Y-chromosome Hgs consist of Y chromosomes with the same set of mutations inherited from a common paternal ancestor. These haplogroups can be organized in a phylogenetic tree representing the evolutionary history of paternal lines. Technological advancements increased during time the number of SNPs and brought to the refinement of the tree [73, 77, 78].

4.4.2 Mitochondrial DNA

Mitochondria are one of the defining features of eukaryotes and the adoption of this organelle was a fundamental step in evolutionary history [79]. The relationship between the cell and the mitochondria also determined the aspects of replication and transmission of mtDNA, the small size, high gene density and copy number variation of the mitochondrial genome [80]. The human mtDNA is a circular, double-stranded DNA molecule (Figure 8).

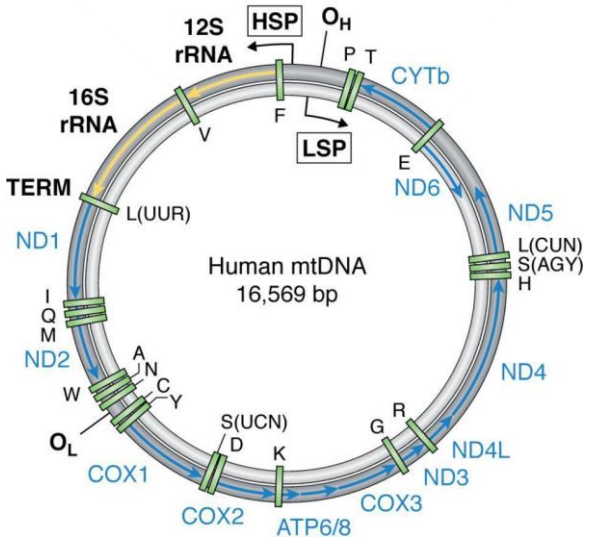


Figure 8. Schematic representation of the human mtDNA. The heavy strand is in a darker color and the light strand is in a lighter color. Genes are represented in blue, while rRNAs are colored in yellow and tRNAs in green (and indicated by the one-letter name of the corresponding amino acid). Origins of replication of the L-strand (O_L) and H-strand (O_H) and the promoters for transcription for the two strands (HSP and LSP) are reported. Modified from Figure 1 of Basu et al. [78] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The two strands are defined as H (heavy) and L (light), the first is guanine-rich while the L-strand is cytosine-rich. The number of mitochondria per cell and of mtDNA molecules per mitochondrion varies greatly, according to the tissue and the cell type, but it is generally relatively high. The first human mitochondrial genome, that became the reference mtDNA sequence (Cambridge Reference Sequence, CRS), was published in 1981 [81], and then revised later in what is nowadays referred to as the revised CRS (rCRS) [82]. It is 16,569 bps in length, and all gene positions are referred to this reference molecule. The human mtDNA contains 37 genes: 13 encode for proteins, 2 for ribosomal RNAs (rRNA) and 22 for transfer RNAs (tRNA). The coding region of the mtDNA constitutes most of the genome, while the main non-coding region is referred to as the “control region”, or D-loop (displacement loop), which is involved in the control of transcription and replication [83]. The D-loop is 1,121 bps in length (nucleotide positions, nps, 16024-576) and it is the most polymorphic region of the mtDNA, particularly in three so-called hypervariable segments: HVS-I (nps 16024-16400), HVS-II (nps 44-340), HVS-III (nps 438-576). As previously stated, the mtDNA is maternally inherited and does not undergo recombination [84–86]. Moreover, this molecule evolves at a higher rate than the nuclear genome, with the D-loop having a mutation rate even higher (especially in the HVS regions) [87]. These features, together with the high copy number per cell, made it a crucial tool for evolutionary genetic studies in the 1980s and 1990s. The mtDNA

phylogeny was built starting from restriction fragment length polymorphisms (RFLPs) and the most common haplogroups were named using alphabetic labels [88], starting from the 1990s. A-G letters were given to American and Asian lineages [88, 89], H-K to Europe [90], and L to Africa [91]. The current mtDNA phylogeny and haplogroup nomenclature (http://www.phylotree.org/) is the result of a series of studies on whole mtDNAs covering an increasing number of individuals from populations across the world. Very recently, the current worldwide phylogeny based on PhyloTree 17 [69] has been updated, including a revised phylogenetic tree with 966 new haplogroups [92]. This new version is called PhyloTree 17 Forensic Update and it has been implemented in the EMPOP database (https://empop.online).

All mtDNA population genetic studies associating sequence variation to time models make assumptions about the molecular clock. The vertebrate mtDNA mutation rate is not uniform across the mitochondrial genomes and its estimate is complicated by other factors. For example, heteroplasmies (the existence of different mtDNA types in the same individual) and germline bottlenecks leading to the loss of many *de novo* mutations [93], or directionality of mutations [94] and higher effective transition/transversion rate biases [95] are all factors that must be considered. Overall, human mtDNA mutation rate is over an order of magnitude higher than the nuclear rate. The molecular clock models and methods most used in humans are based on the two most recent estimates of the mtDNA mutation rate. The first is $2.33 \pm 0.2 \times 10^{-8}$ substitutions per nucleotide per year, published by Soares et al. [55], and corrects for the effect of selection. The second one, $2.70 \pm 0.2 \times 10^{-8}$ substitutions per nucleotide per year, published by Posth et al. [96], is based on the divergence between ancient and modern human mitogenomes. Estimates of the mtDNA time to the MRCA vary depending on the dating technique and mutation rate but are in the range of 200-100 thousand years ago (kya) for the ancestral mtDNA molecule [71], that is also called “mitochondrial Eve”.

The root of the mtDNA phylogeny and the most diverse branches are restricted to African populations (Figure 9).

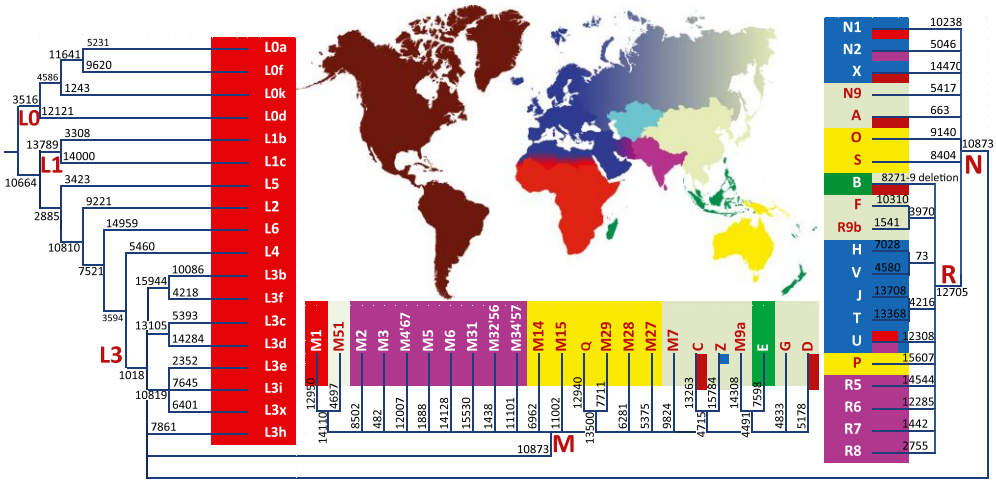


Figure 9. Representation of the mtDNA worldwide Hg distribution. Labels are reported according to [69]. Only a single branch-defining mutation, preferably from the coding region, is shown. The main geographic features of haplogroup distribution are highlighted with colors. Adapted from Figure 2 of Kivisild [71] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The first seven bifurcations in the tree separate the strictly sub-Saharan African branches (L0-L6) from those (as L3) that are shared between African and non-African populations. Outside Africa, haplogroups L0-L6 are rare and restricted to areas where migrations from Africa have occurred, as Mediterranean Europe, West Asia, and the Americas [97]. Virtually every non-African mtDNA lineage derives from Hg L3 only (Figure 9), indicating a strong bottleneck of mtDNA diversity at the time of the out of Africa dispersal [98]. The separation of the two macro-Hgs M and N from the African L3 clade has been dated to 95-62 kya [99], while the internal coalescence time of these two clades has been estimated from 70-40 kya [100]. Archaeological evidence [101] and the mtDNA of Ust-Ishim (Western Siberia) skeletal remains, dated at 45 kya and whose mitogenome falls at the root of Hg R, seem to confirm this estimate [102]. Hgs M and N are distributed in Eurasia, Australia, Oceania, and the Americas, but each of their sub-clades has a more specific regional configuration. Eurasian lineages, as U, HV, JT, N1, N2, and X, are present today in Europe, South-West Asia and North Africa [103]. Hgs R5-R8, M2-M6, and M4'67 are virtually restricted to East Asia [104]. In the Americas, the last continent to be colonized by anatomically modern humans, the mtDNA variation is represented primarily by Hgs A to D and a specific sub-clade of Hg X (X2a) [105]. Based on complete mitogenome analyses, at least 16 Indigenous American founding lineages have been defined so far [106–108, 105, 109–111]. These clades are distributed differently in the Americas, with some Hgs spread all over the double continent (called "pan-American"), and others restricted to specific regions. Moreover, based on mtDNA data, migrations in the Americas have been associated with at least three distinct demographic events, as also proposed

in the tripartite model [112]. The first migration wave into the Americas was dated at 18-15 kya involving lineages (A2, B2, C1b, C1c, C1d*, C1d1, D1, D4h3a, and D4e1c) found in both American continents, and for this reason called “pan-American”. The second event was the migration of people carrying C4c and X2a lineages to the eastern region of North America. The third event was the spread of Palaeo-Inuit D2a lineages ~5 kya across the Arctic through Northern Canada and Greenland, which were replaced, in the same region, by the spread of Thule-associated people carrying A2a, A2b, and D3 lineages. However, the association between haplogroup A2a and an ancestral Palaeo-Inuit route, inferred by the geographic distribution of modern mitogenomes [106] has been recently questioned due to scarce ancient DNA (aDNA) evidence associating remains of Palaeo-Inuit cultures, as Saqqaq and Dorset, with haplogroup D2a [113]. Several Hgs have been associated with two main events in the peopling of Oceania. Lineages M14, M15, M27-M29, Q, P, O, and S, observed only in Australia and Melanesia, are related to the first colonization of Papua New Guinea and Australia 47-43 kya. B4a1a1 clades, instead, are related to a second more recent migration associated with the diffusion of Austronesian languages ~7 kya in Southern Australia [114, 115].

All these results, however, were not sufficient to define some issues as, for example, the source and number of Mesolithic and Neolithic gene flows in Europe, the colonization, and the subsequent migrations, within the Americas, and the first peopling of Oceania. All these issues have been subsequently (and currently) addressed by studies on autosomal markers and on ancient DNA, either answering these questions, or revealing even more complex scenarios, both at a macro- and microgeographic scale.

4.5 Autosomal markers

Uniparental systems can be thought of as two *loci* that can be used to reconstruct the paternal and maternal history of an individual or a population. However, they describe only two ancestral paths out of the thousands that contributed to modern populations. This complexity can be grasped by using autosomal markers. The haploid human genome is ~3.2 Gb, organized in 23 separate molecules, referred to as the chromosomes. In a human individual, there will be a total of 46 paired chromosomes in somatic cells, and 23 chromosomes in gametes. Each member of a pair is inherited from a parent. 22 chromosomes are called autosomes, while the remaining are called sex chromosomes, X and Y, the latter occurring only in males. Autosomes undergo meiotic recombination [116] and the genetic material is shuffled at every generation. As described above, while variants on different chromosomes are unlinked, variants on close *loci* on the same chromosome are linked together and can be co-inherited from the same parent as

haplotypes. These haplotypes can be broken by recombination. However, in the same genome recombination frequency can vary greatly, with region where it does not occur, and regions, also called *hotspots*, in which recombination frequency is much higher. Highly recombining segments are small (1-2 Kb) and are separated by larger regions with a low recombination activity [116]. These differences across the genome must be considered when performing population genetic studies based on haplotypes. Variability on autosomal markers can be investigated through whole-genome sequencing or genotyping.

4.6 Ancient DNA

Since the mid-2000s, advancements and broad applications of high-throughput sequencing (HTS) techniques enabled rapid and cost-effective sequencing of genomes and paved the way for a genomic era also in the field of ancient DNA (DNA obtained from sub-fossil remains of past organisms) [117]. This archaeogenomic [118] revolution started in 2010 with the sequencing of the first ancient human genome [51] and of the first Neanderthal genome [35], and expanded almost exponentially. Nowadays, thousands of ancient human genomes have been sequenced (Figure 10), and hundreds of ancient genomes of other species and even microorganisms have been published [119–124]. Very recently, the time limit of aDNA studies has been increased with the sequencing of a one-million-year-old mammoth [125].

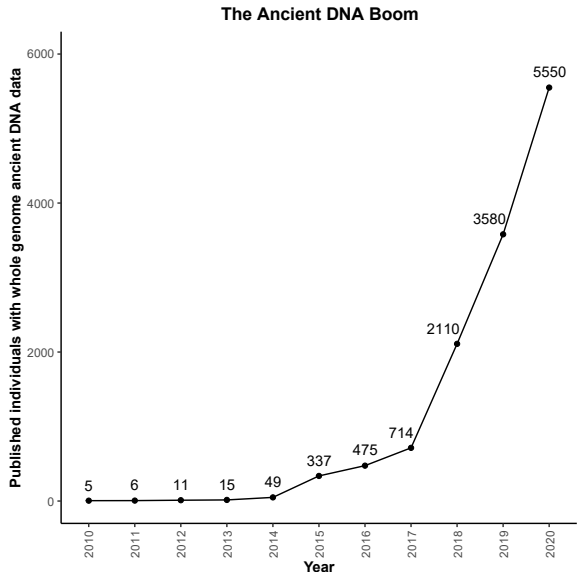


Figure 10. Graph representing the number of ancient genomes that have been published since 2010 and until 2020 (modified from https://reich.hms.harvard.edu/research).

Already in the 1980s, it was shown that DNA can survive long after an organism's death [126], but it is inherently highly degraded and often chemically altered. There are different challenges that must be faced when working with ancient DNA, such as sequence degradation, post-mortem damage, and modern DNA contamination. After the death of an individual, DNA starts decomposing. This results into the fragmentation of the DNA strands into shorter fragments (~40-100 bps), with a consequent reduction of quantity and quality (Figure 11A) [127–130]. Cytosine deamination at the ends of each DNA fragment results into an uracil which is read as thymine when sequencing, mimicking a true polymorphism. This will determine the increase in number of the C->T variants scored toward the 5' end of the sequence fragments and a complementary increase in G->A variants at the 3' end (Figure 11B). It is important to note that some strategies to sequence aDNA may produce different patterns [131, 132] and that there are methods to repair these changes before sequencing [133, 134].

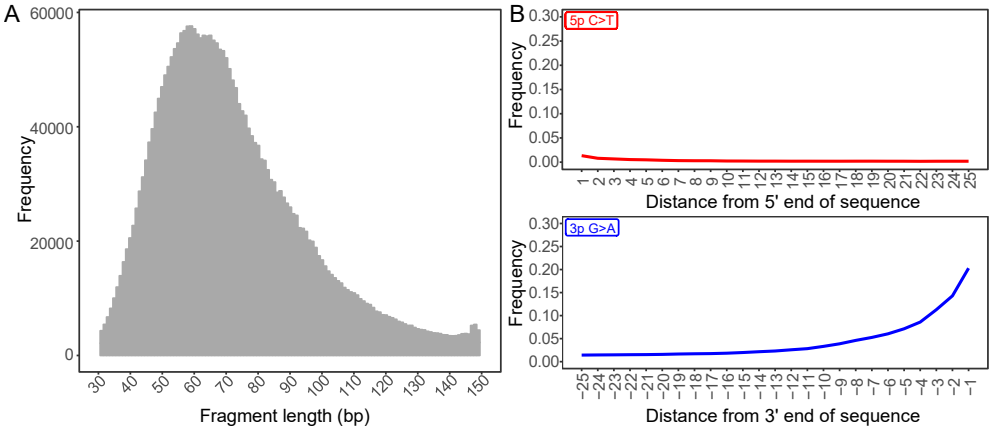


Figure 11. Characteristic damage patterns of an ancient low-coverage genome produced in our aDNA laboratory: A) Length distribution of sequenced fragments; B) Enrichment of deamination toward the fragments ends. In this example, the deamination pattern is present at the 3' end but less evident at the 5' end, due to the AT-overhangs library technique [132, 135].

These aDNA features are correlated with environmental conditions and time. In the same environment, the older the remains are, the stronger the observed post-mortem damage will be. Moreover, in environments that are more humid and warmer (e.g., in the tropics) *post-mortem* modifications occur more rapidly, heavily impacting DNA preservation [136–138]. On the other hand, thanks to the development of specific statistical analyses, these aDNA features have become useful to distinguish real ancient DNA from modern contaminants. In fact, the other great challenge of aDNA analyses is modern DNA contamination, that can originate from modern human DNA or from exogenous DNA of soil bacteria and other microorganisms [129, 139]. Human handling implies a risk of contamination during excavation, storage, and processing of ancient remains. The contaminant DNA would typically be

younger, of higher quality, and at a greater concentration, if compared to the authentic endogenous DNA. For non-human/hominid remains, contamination can be handled by mapping to reference genomes (species-specific and contaminant-specific). For ancient human DNAs, modern human contamination is particularly challenging. Clean lab procedures and dedicated facilities have substantially reduced the risk of contamination. However, computational methods are required to assess and estimate the extent of contamination in sequence data, and such approaches are part of the standard quality checks performed in human aDNA studies. As previously stated, some of these methods rely on the peculiar features of aDNA, such as cytosine deamination [140]. In recent years, aDNA studies have led to some major breakthroughs in our understanding of the evolution and prehistory of our species. Paradigmatic examples are the discovery of a new hominid, the Denisovan, the introgression of Neanderthal sequences in the genome of anatomically modern humans [35], the identification of ancestral components in Eurasia and in the Americas [5, 141, 142], and the possibility to recalibrate molecular clocks using radiocarbon-dated samples [96]. Finally, recent applications of aDNA studies have involved metagenomic approaches to analyze pathogens and sedimentary aDNA, in which researchers try to extract and sequence aDNA to find which species were present in soil and stratifications of archaeological sites [143, 144], as well as palaeoepigenomic applications focusing on generating epigenomic data from ancient organisms [145].

4.7 Overview of human history and migrations

4.7.1 The origins of anatomically modern humans

As mentioned before, mtDNA studies showed that anatomically modern humans (AMH) evolved in Africa and then migrated to the rest of the world [146–149]. This *out of Africa* model proposes a single transition from archaic hominins to AMH in Africa followed by later migration(s) outside Africa, replacing other extant hominin populations. This model, now widely accepted and supported by genetic data [38, 96, 150, 151], has been in contrast with the so-called *multiregional hypothesis* that in its various propositions suggested a transition from *Homo erectus* to AMH in many regions of the world [152] or an independent origin of AMH from groups of *Homo erectus* in different regions worldwide around one million of years ago [153, 154].

4.7.2 Archaic and modern humans' interactions

Our closest extinct hominin species were the Neanderthals and the Denisovans, both of which disappeared ~50-40 kya. The first genomes of Neanderthals and Denisovans have been sequenced in 2010 [35, 155]. The analysis of these genomes showed that they are more closely related to each other than to AMH [155]. Archaic humans are estimated to have diverged from AMH ~550 kya, then splitting from each other ~400 kya [156–158]. It has been shown that genetic admixture between archaic humans and AMH occurred after the *out of Africa*, and it has been detected as recently as ~40 kya [159, 160]. These signals are present in modern non-Africans, even though with varying admixture proportions and geographical distributions [35, 155, 158, 161–163]. Admixture with Neanderthals probably occurred soon after the *out of Africa* (~55 kya) in Southwestern Asia. AMH also admixed with Denisovans in Eastern Asia, as shown by the presence of Denisovan ancestry in present-day East Asians and Oceanian populations [131, 155, 164], reaching the highest values in Philippine Ayta [165]. Recently, a ~50-ky-old Denisovan individual has been shown to have had a Neanderthal mother and a Denisovan father [166], indicating that past archaic humans mixed also with each other.

4.7.3 Africa

Many locations in Africa have been proposed for the origin of AMH, considering archeological records as well as estimates on genetic diversity and divergence time [167–171]. These estimates are ~250-200 kya [168] or ~360-260 kya for the divergence of AMH in Africa [172], leading to the formation of deeply diverged African populations. This is highlighted by the greater amount of genetic diversity that can be found between any pairs of African genomes than that found between any non-African genomes [173, 174]. Most of all non-African human ancestry stems from a worldwide expansion out of Africa that started ~60 kya [39]. This dispersal was accompanied by a bottleneck that left a strong signature on the genetic variation of all non-African populations [175]. After the *out of Africa*, AMH populations expanded across the world (Figure 12), with multiple founder effects and bottlenecks, thus further explaining the lower genetic variation compared to African populations [174, 176].

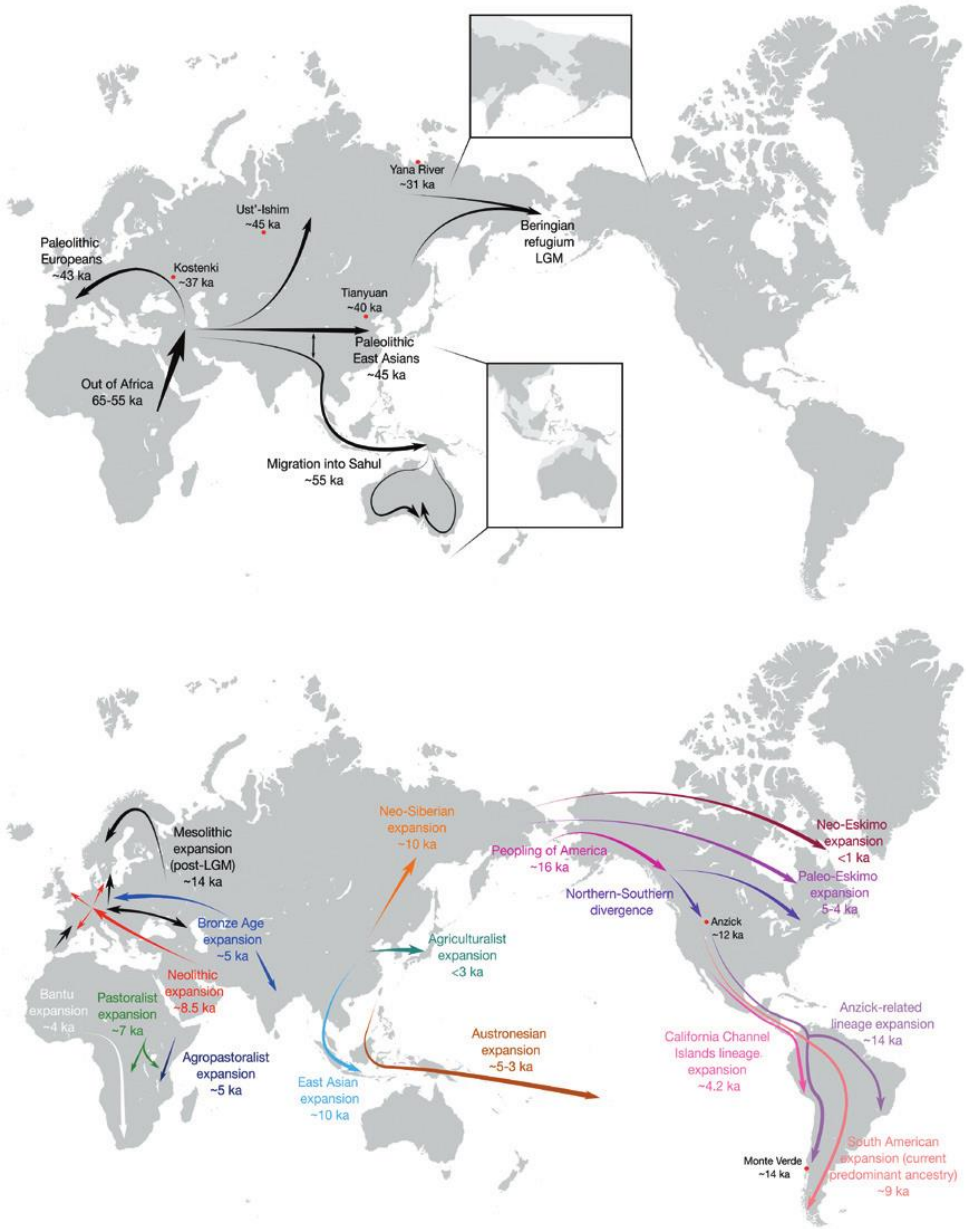


Figure 12. Major human migrations across the world inferred through analyses of ancient genomic data. Some migration routes remain under debate, such as those used to populate the Americas. In the above map, early movements of modern humans after they left Africa are highlighted. In the below map, movements after the Last Glacial Maximum are reported. Modified from Figures 1 and 2 of Llamas et al. [177] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The reconstruction of human history in Africa has just begun due to the poor preservation of aDNA at those latitudes. Ancient genomes revealed that

some early Africans were genetically connected with populations in the Near East [169]. More recently, between 5-1 kya, multiple admixture events among herders and farmers in sub-Saharan Africa gave rise to present-day East African populations [168, 169, 178, 179]. Due to these complex events and the scarce availability of ancient human genomes, population structure and movements within Africa remain debated [117].

4.7.4 Asia

Most evidence indicates that Asia was colonized through at least two early waves of migration. One wave included the ancestors of Australians and Papuans while the other included the ancestors of east Asians, with admixture between the two [180]. Other evidence suggests that there was only one dispersal event [38]. In east Asia, an ancestry related to a 40 ky-old [99] and a ~33 ky-old individuals from present-day China was widely distributed across northern east Asia before the last glacial maximum [181, 182]. Two ~31.6 ky-old genomes from Yana RHS site in Northeastern Siberia [183] reflect a population that was ancestral to the 24 ky-old genome from the Mal'Ta-Buret' culture from southern central Siberia [113] and the ~18 ky-old Afontova Gora genome [184], which show strong connection to both western Eurasians and Indigenous Americans, but weaker affiliations with east Asians and Siberians. This may be due to different geographic distribution of genetic signatures during the Upper Paleolithic. Another genome from the 38-36 ky-old Kostenki 14 individual [185] shows a close affinity to contemporary western Eurasians but not to east Asians. After the LGM, a potential population shift occurred related to a ~19 kyr-old individual from the Amur region in China [181]. Population changes also occurred in Siberia, where populations carrying the so-called Ancient North Siberians and Ancient North Eurasians ancestries were replaced by Ancient Palaeo-Siberians, represented by a 9.8 ky-old skeleton from Kolyma River [183]. Around 4.9 kya, steppe populations also expanded eastward, contributing to individuals associated with the Afanasievo culture in the Altai region [141, 186]. Subsequently, between 2.5 and 3.5 kya, Yamnaya people in central Asia were locally replaced by individuals from the Sintashta culture [186], who came from the Urals and Europe and admixed with east Asians.

4.7.5 Oceania

Archaeological evidence suggests that human populations from Southeast Asia have settled in Sahul (present-day Australia, Tasmania, and Papua New Guinea) before ~50 kya [187], probably through a single migration wave [115, 188]. In the Oceanian archipelagos multiple waves of migration and

admixture events occurred within the past several thousand years. About 3.2 kya the Lapita culture expanded into remote Oceania, carrying mostly Austronesian-related ancestry [189]. However, this ancestry was nearly completely replaced by the Papuan-related one ~2.7-2.3 kya in the most western islands of Oceania [190, 191]. Additional migrations from Philippines have been proposed [192], highlighting the complexity of the peopling of Oceania. An important matter of debate nowadays is whether there has been contact/admixture between Polynesians and Indigenous Americans. Recent analyses of present-day genomes suggest a contact ~800 years ago [193], while another study did not support this hypothesis [191].

4.7.6 Europe

Genetic data from early modern humans that lived in Eurasia allowed to identify several early modern human lineages [194]. Some of these lineages, as those represented by individuals older than 40 ky from present-day Russia [102], Romania [159], and Czech Republic [195], did not contribute to the current populations. Others, dated to ~40-24 kya are genetically connected to modern populations [183–185]. Ancient Europeans whose ancestry is found in present-day west Eurasians are represented by a ~35-ky-old individual from Belgium [184]. Moreover, west Eurasians diverged ~39 kya from Ancient North Siberians and ~46-43-ky-old genomes from Bacho Kiro Cave in Bulgaria [160] connect ancient genomes from present-day Belgium and China [160, 182]. All these complex dynamics testify for the contemporary expansion and transition of initial Upper Paleolithic cultures [196]. The exact contribution from early Europeans is still debated. There is evidence of turnover in the genetic composition of Europeans before the LGM, followed by the dispersion of western European hunter-gatherers [184] out of glacial refugia after the LGM ~15 kya. Europe then witnessed the spread of farming, that started during the Neolithic ~8.5 kya. However, it was debated if agriculture was acquired in Europe through migration of farming populations or through the transmission of ideas and culture. Ancient DNA analyses confirmed that Neolithic farming populations from Near East largely expanded throughout Europe and admixed in the subsequent millennia with Mesolithic hunter-gatherers, demonstrating that farming was introduced by the migration of people [197]. The Neolithic lifestyle helped to increase the size of populations [198]. Around 4.9 kya, during the early Bronze Age, the so-called steppe ancestry, closely related to individuals associated with the Yamnaya culture from the Pontic-Caspian-Ural steppe region, expanded westward and eastward [141]. This ancestry appeared in central Europe and gave rise to the population associated with the Corded Ware culture [141]. This migration was probably linked to conquests and technological innovations such as horseback riding, and it was also suggested that it brought Indo-European languages into Europe. Around 4.6 kya, individuals

with steppe ancestry arrived in the British Isles, coinciding with the spread of the Bell Beaker Complex, replacing almost completely the local gene pool within a few hundred years [199]. Thus, present-day European ancestry consists of three major genetic components, which reflect: i) the contributions of hunter-gatherers to the recolonization of Europe after the LGM, ii) the migration of Neolithic farmers from Anatolia, and iii) the late-Neolithic period and Bronze Age migration from the east.

4.7.7 The Americas

The first peopling of the Americas occurred from Siberia, through Beringia, possibly ~17.5-14.6 kya, although this peopling may have started before, during or immediately after the LGM (~26-18 kya) [200, 201]. This may predate the first entrance into the Americas, but further studies are needed to confirm this hypothesis. Initial settlement attempts were followed by a more widespread peopling that reached southern South America as early as ~15 kya [202]. Crucial early information about the genetic history of America was based on uniparental systems [107, 106, 203–206, 109, 110], but the emergence of modern and ancient genomic data is improving the knowledge about the first settlement and the following migrations [175, 197, 207] which were revealed to be far more complex than previously thought (Figure 13).

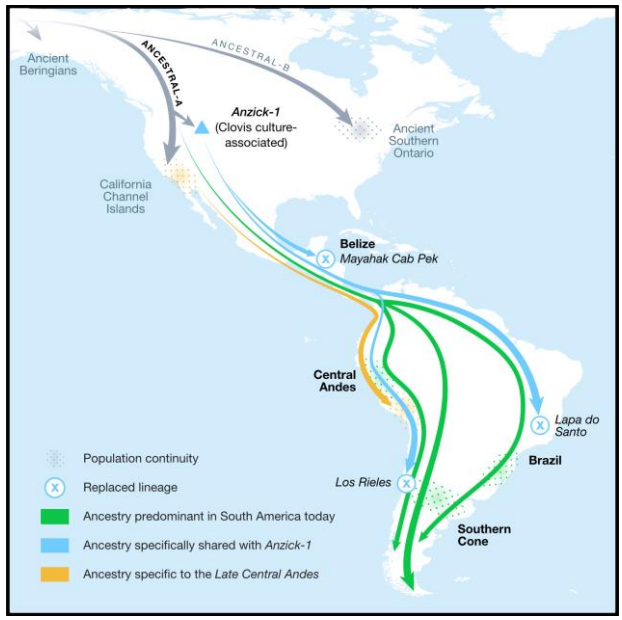


Figure 13. An example of the complexity of the first peopling of the Americas and subsequent migration, admixture, and replacement events, with a particular focus on Central and South America dynamics. Graphical abstract from Posth et al. [208] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Genomes from the Yana RHS (~31.6 kya) and the Mal'Ta (~24 kya) sites show that Siberia was occupied by a population defined as Ancient North Siberian (ANS) [183]. The admixture between ANS and an east Asian population gave rise to at least two distinct lineages [183, 209, 210]. One diverged as the basal American branch [142, 183, 211], whose origin and divergence time are still uncertain [207]. Some evidence suggest that it was isolated, either in western Beringia or further to the south [210, 211]. An isolation during the LGM is compatible with the Beringian standstill model [111, 209, 212, 213], which suggests that dispersal into the Americas occurred after a long pause in Beringia, during which several lineages emerged [142, 211]. One of them characterized the Ancestral Native Americans (ANA), which then split first into the Big Bar lineage 21-16 kya, and then into Northern Native American (NNA) and Southern Native American (SNA) populations [214, 215, 209, 216, 217]. The location of this last split, either in Beringia or south of the continental ice sheets is still debated. NNAs remained in northern north America and then possibly moved further northward [142, 211, 217]. SNAs rapidly spread southward, reaching southern south America as early as ~15 kya [202], as evidenced by the Monte Verde archaeological site and by the genetic closeness of SNA ancient individuals dated to the same period (~10 kya) but excavated both in North and South America [142, 208]. It is suggested that there have been at least two migrations during the Late Pleistocene of SNA groups into South America, given the different genetic affinities of ancient individuals from Argentina, Brazil, and Chile to the SNA Anzick-1 Clovis child (~12.8 kya) excavated in Montana [208, 218]. This large-scale migration, together with the relatively small size of SNA population [212], increased the possibility of isolation and divergence, evidenced by many splitting within the SNA groups while moving south [142], which led in turn to the considerable variation in ancient South Americans [208, 215, 219]. More recently, during the Holocene period, there was a complex movement of people both into and within the Americas. One example are the migrations of Palaeo-Inuit first and then Thule people in northern North America [51, 113, 197]. Further south, there was a Mesoamerican-related expansion toward North and South America [142]. This gene flow, including traces of an unsampled population A (UPopA) [142] is evident in different proportions especially in South American populations. A second SNA lineage spread into South America ~5 kya, largely replacing earlier groups with close affinities to Anzick-1 [142, 208, 218]. During the Holocene, population movements also occurred in the Caribbean [220, 221]. A first wave of migration from South America ("Archaic Caribbean") was then replaced by a second one ("Ceramic Caribbean"), possibly related with Arawakan-speaking groups from the Amazonian region, although this connection is still debated [220]. Some genetic studies have revealed a genetic affinity between Amazonian and Australo-Melanesian populations [209, 222]. Some signals were found in present-day people of Rapa-Nui, but these were not confirmed [223, 224]. A recent study found Polynesian individuals with Indigenous American admixture [193]. Since a trans-Pacific

migration wave seems unlikely, it has been suggested that Indigenous American ancestors might have genetic similarities with an eastern Asian population(s) also related to modern Australo-Melanesians [182, 222], still detectable in some isolated Amazonian populations [182]. This component may have been brought into America by a ghost lineage named Population Y (UPopY) and was already present ~10 kya in Brazil.

5. Aims

As clearly stated in the title, the main objective of this thesis was to unveil the genomic history of human populations in macro- and microgeographic contexts. To achieve this goal, different scientific approaches have been employed. The sequence variation of the maternally inherited mtDNA, a well-established molecular tool in population and forensic genetics, has been investigated in various populations from Eurasia (Umbria in central Italy and Mongolia) and the Americas (Mexico and the Isthmus of Panama). These studies took advantage of the highly defined and well-calibrated mitochondrial phylogenies, in one instance also paralleled by the paternally inherited Y chromosome. In other contexts, mtDNA analyses have been coupled to, or enriched with, the analysis of genome-wide data (Panama and Peru) to achieve higher molecular resolution and statistical power. It is worth mentioning that the genomic screening, particularly of ancient DNA, is the most useful when dealing with Indigenous American populations, since most of their original variants have been lost after European contact. Therefore, in the peculiar Isthmian context, a diachronic comparison between modern genome-wide data and ancient genomes (the first ones from the Isthmus of Panama) has been performed to fulfill the requirements of an archaeogenomic study.

6. Methods

6.1 Modern DNA

6.1.1 DNA collection and extraction

All the DNA samples analyzed in this thesis have been collected by our research group or our collaborators during different sampling campaigns. All participants signed an informed consent form and were asked to provide genealogical information, spoken language, as well as their possible affiliation to an ethnic group (as in the case of Panamanian Indigenous individuals). Genomic DNA was extracted from saliva, collected through buccal swabs and/or mouthwash rinsing. DNA extraction was performed either with the classic phenol/chloroform method, available commercial kits, or automated extraction with the Maxwell® RSC Instrument (Promega). Extracted DNA was quantified with a fluorometric instrument, as the Promega Quantus™ or the Invitrogen Qubit™, and then stored at -20°C.

6.1.2 Modern mitochondrial DNA

A total of 5,839 mtDNA sequences have been analyzed in the works presented in this thesis. In particular, 5,417 control-region sequences and 418 complete mtDNAs have been obtained as described hereafter.

6.1.2.1 Amplification and sequencing

Two different PCR approaches were used based on the molecular target. Standard PCR was employed to amplify the mtDNA control region (D-loop), a ~1.1 kilo base pairs (kb) long fragment, using specific flanking primers [4]. The complete mtDNA, instead, was amplified with two long-range (LR) PCR reactions for each molecule. This protocol makes use of two primer pairs to amplify two partially overlapping fragments of ~8-9 kb in length each [1, 5, 108]. LR and standard PCRs were evaluated on 1% and 2% agarose gels, respectively.

The amplified control regions were then enzymatically purified using the ThermoFisher ExoSAP-IT™ system, which utilizes two hydrolytic enzymes to

remove unwanted nucleotides and primers from PCR products. The Exonuclease I removes residual single-stranded primers and any other single-stranded DNA, while the Shrimp Alkaline Phosphatase removes the unincorporated nucleotides from the PCR mixture. The control regions were sequenced at the BMR Genomics (https://www.bmr-genomics.it/) through Sanger sequencing using the Brilliant Dye terminator 1.1 kit. All D-loops were sequenced using a forward primer [4], while a reverse primer [4] was also used for those sequences harboring the transition T16189C, which results in a poly-C tail causing premature termination of the sequencing reaction. The raw data were provided as .ab1 (Applied Biosystems) files, which contain an electropherogram and the DNA sequence. Sanger sequencing was also used to cover small sequence gaps in complete mtDNAs sequenced with NGS and also, in a few instances, to sequence the entire mitogenome. In this latter case, an established protocol for amplifying and sequencing overlapping fragments across the complete mtDNA was used [225, 226].

The two partially overlapping LR amplicons covering the whole mtDNA molecule were combined, considering the relative amount of each fragment (based on the band intensity on gel electrophoresis) and their relative length. The PCR mixture was then purified either with the Promega Wizard® SV Gel and PCR Clean-Up System, or with the Geneaid Presto™ 96 Well PCR Cleanup Kit, following the manufacturer's protocols. The complete mtDNAs were sequenced through an NGS approach [5]. Libraries were prepared using the Nextera XT DNA Library Preparation Kit, which uses tagmentation and reduced-cycle PCR amplification to fragment DNA and add adapters and barcode indexes for multiplexing. The Nextera XT Index Kits, which offer up to 384 unique index combinations, were employed, enabling accurate assignment of reads and efficient use of the flow cell. Libraries and final pools were checked on an Agilent Tapestation 4150 system, an automated capillary electrophoresis platform for nucleic acids. Final libraries were then sequenced on an Illumina MiSeq at the Genomic and Post-Genomic Unit, IRCCS Mondino Foundation in Pavia. The run was performed using the MiSeq Reagent Kit v2 (300-cycles, paired-end reads), with either the standard, micro, or nano flow cell, obtaining a maximum output of 5.1 Gb, 1.2 Gb, and 0.3 Gb, respectively, for up to 15 million reads per run.

6.1.2.2 Sequence analysis

After Sanger sequencing, the .ab1 files, containing the electropherograms, were aligned and compared to the rCRS reference sequence [82] using Sequencher v4.9 (http://www.genecodes.com/), which was also used to manually determine the control-region haplotypes (the complete set of variants in a mtDNA sequence). These were used for haplogroup classification through the software HaploGrep v2.1.21 [227]. This

classification is based PhyloTree 17 [69], the mtDNA tree estimated from worldwide data.

The Illumina MiSeq system generated raw data files in binary base call (BCL) format. The BCL files were converted into FASTQ files using the Illumina software bcl2fastq v2.19.0.316. This program assigns each read to a single sample, based on the unique dual combination of index sequences (demultiplexing). The two FASTQ files for each mtDNA sequence were then analyzed using a Unix-based pipeline developed by our research group. Here, I report the last version of this pipeline, which I contributed to update during my PhD. The very first step is a quality check of the raw reads. This is achieved using the software fastQC v0.11.8 [228], that produces a quality control report consisting of different modules to identify potential problems in the sequence data.

fastqc --threads 8 -f fastq ${sample_name}_*.fastq.gz -o ${fastqc_raw}

After the quality check, the first step of the pipeline is the quality and adapter trimming of raw FASTQ. The software trim_galore v0.6.4 [229] is used.

trim_galore --phred33 --paired -q 30 --nextera ${sample_name}_*_R1_*.fastq.gz ${sample_name}_*_R2_*.fastq.gz

The flag *--nextera* tells the program which specific Illumina adapters have been used and are therefore to be trimmed. The *-q 30* flag instead is used for trimming low-quality ends from the reads in addition to adapter removal. The value 30 is the minimum quality threshold in Phred scores (*--phred33* instructs the software to use ASCII+33 quality scores as Phred scores), which indicate the probability of a correct base calling for each base. After trimming, the FASTQ files are again quality checked using fastQC.

fastqc --threads 8 -f fastq ${trm_fastq}/${sample_name}_*.fq.gz -o ${fastqc_trm}

Trimmed reads are then aligned to the rCRS mitochondrial reference sequence using the software Burrows-Wheeler Aligner (BWA) v0.7.17 [230], with the algorithm MEM that is specific for long reads (>70 bp), generating a Sequence Alignment Map (SAM) file. The flag *-R* is used to add to the header of the SAM file information on library and sequencing platform. The SAM file is converted into a Binary Alignment Map (BAM), which stores the same information but in less space, by compressing it via bgzip. For this conversion, the software SAMtools v1.9 [231] is used, which contemporarily filters the BAM file by removing all reads that (i) have mapping quality less than 30 (in Phred score; *-q 30*), (ii) are not mapped, (iii) are not a primary alignment, (iv) fail platform/vendor quality checks, (v) are supplementary alignments (*-F 2820*), and by keeping only those reads that are mapped in proper pairs (*-f 2*).

```
bwa mem \
-R "@RG\tID:${sample_name}\tSM:${sample_name}\tLB:nextera\tPL:Illumina" \
-t 8 \
${mtDNA} \
${trm_fastq}/${sample_name}_*_R1_*.fq.gz
${trm_fastq}/${sample_name}_*_R2_*.fq.gz \
| samtools view -q 30 -F 2820 -f 2 -bho ${bam}/${sample_name}.bam
```

SAMtools is also used to sort the alignment file by coordinates.

```
samtools sort ${bam}/${sample_name}.bam \
-o ${bam}/${sample_name}.sortco.bam
```

The last important filtering step is the removal of duplicate reads that derive from multiple PCR amplifications of the same DNA fragment and that may cause error propagation if not removed. To this end, duplicates are removed using the tool *MarkDuplicates* v2.21.6 from the Picard software package (broadinstitute.github.io/picard). This tool works by comparing sequences in the 5' positions of both reads and read pairs.

```
picard MarkDuplicates \
I= ${bam}/${sample_name}.sortco.bam \
O= ${bam}/${sample_name}.sortco.rmdup.bam \
REMOVE_DUPLICATES=TRUE \
METRICS_FILE= ${rmdup}/${sample_name}_picard_metrics.txt
```

Final BAM files are indexed using SAMtools.

```
samtools index ${bam}/${sample_name}.sortco.rmdup.bam
```

We then use a combination of SAMtools *depth* command and *awk* to retrieve the possible gap intervals for each sample in a table, meaning all positions compared to the reference whose depth is less than a specific value. In our analyses, this depth threshold is usually set to 3 or 5. This step allows us to identify possible gaps which must be controlled manually on the alignment and/or covered with Sanger sequencing.

```
samtools depth -a ${bam}/${sample_name}.sortco.rmdup.bam \
| awk '$3 < 5' > ${depth_less5}/${sample_name}.less5.txt
```

```
awk 'function output() {print start (prev == start ? "" : "-"prev) } NR == 1 {start =
prev = $2; next} $2 > prev+1 {output(); start = $2} {prev = $2} END {output()}'
${depth_less5}/${sample_name}.less5.txt \
| transpose -t | sed -e 's$\t$; $g' > ${depth_less5}/${sample_name}.gaps.txt
```

We also perform quality control of the alignment sequencing data and a summary of its main features (such as the average depth of coverage) using Qualimap v.2.2.2 [232]. The results are outputted in both a PDF and HTML format.

qualimap bamqc -sd -sdmode 2 -bam ${bam}/${sample_name}.sortco.rmdup.bam -outdir ${qualimap} -outformat PDF:HTML

The last step is the variant calling. The Genome Analysis ToolKit (GATK) v4.2.2.0 [233], a platform with tools to analyze NGS data developed by the Broad Institute (https://www.broadinstitute.org/), is used. The tool HaplotypeCaller calls SNPs and indels via local re-assembly of haplotypes in active regions. During variant calling some filters are applied for a variant to be called. The flag *-mbq 20* sets to 20 the minimum base quality required to consider a base for calling, while the flag *-stand-call-conf 30* is the minimum Phred-scaled confidence threshold at which variants should be called. Finally, *-ploidy 2* sets the analyzed sequences as diploid. This is because, although analyzing haploid mtDNA, we want to detect heteroplasmies, which will be called with a 0/1 genotype. The output of variant calling is the Variant Call Format (VCF).

```
gatk --java-options "-Xmx8g" HaplotypeCaller \
-I ${bam}/${sample_name}.sortco.rmdup.bam \
-O ${vcf}/${sample_name}.vcf.gz \
-R ${mtDNA} \
-mbq 20 \
-ploidy 2 \
-stand-call-conf 30
```

Finally, a set of bash commands produce a summary file for each mtDNA sequence. This file includes information from previous outputs and calculate various statistics, such as the average depth of coverage, the total number of reads, the number of mapped reads and duplicates, and the percentage of the mitogenome with an average depth of coverage greater than 3, 5, and 10.

The VCFs were used for haplogroup classification with the stand-alone version of HaploGrep 2 v2.1.21 (github.com/seppinho/haplogrep-cmd). Mitochondrial haplotypes were also checked by manually inspecting BAM files using SAMtools *tview* command.

6.1.2.3 Genetic diversity

The calculation of mtDNA molecular diversity indices was performed with DnaSP v.6 [234]. In details, the haplotype diversity (Hd) with its standard

deviation, the nucleotide diversity π (pi) and the average number of nucleotide differences (k) were determined. Haplotype diversity represents the probability that two randomly sampled haplotypes are different in a given population [235]. Nucleotide diversity is defined as the average number of nucleotide differences per site in pairwise comparisons among DNA sequences and was estimated by assessing windows of 100 bps with step size of 50 bps centered at the midpoint [235]. The average number of nucleotide differences measures the degree of polymorphism within a population and is defined as the mean number of nucleotide differences per site between two DNA sequences randomly chosen from the sample population [236]. Another index based on haplogroup frequencies, the heterogeneity, was computed using the standard method of Nei [235]. Haplogroup frequencies and distributions were visualized with either Excel, Tableau (https://www.tableau.com/), or R [237], while statistically significant differences were estimated with either the Chi-square test of independence or the Fisher's exact test of independence using R [237] and the XLSTAT add-on for Excel. Genetic pairwise distances between individuals were computed using MEGA X [238], as the proportion of nucleotide sites at which two sequences being compared are different (*p-distance* method). This is obtained by dividing the number of nucleotide differences by the total number of compared nucleotides. Distances were calculated using only variable sites and disregarding indels and private mutations. These distances were also converted into a dissimilarity matrix, using the R functions *mean()* and *xtabs()* [237]. The resulting matrix was used to perform a multidimensional scaling (MDS) using the R function *cmdscale()* [237]. PCAs, used to summarize quantitative multivariate data, were computed on haplogroup frequencies using *prcomp()* from the *stats* R package [237], with the *center* and *scale* arguments set as true. Correspondence Analysis (CA), an extension of PCA suited to explore relationships among qualitative variables, was also performed on haplogroup frequencies using the *CA()* function from the *FactoMineR* R package [239], as in [240].

6.1.2.4 Phylogeny and phylogeography

Before building phylogenetic trees, sequence data were aligned, mostly using Sequencher 4.9, checked and manually corrected to solve indel misalignments. Phylogenetic trees were built using different algorithms. In details, mtPhyl (eltsov.org/mtphyl.aspx), a specific tool for mtDNA analyses, was used to build MP trees. These trees were based on Phylotree 17. Bayesian inferences were obtained using the software Bayesian Evolutionary Analysis by Sampling Trees v.2.6.5 (BEAST) [241]. BEAST 2 can be used as a method for reconstructing phylogenies but also for testing evolutionary hypotheses without conditioning on a single tree topology. Each tree is weighted proportionally to its posterior probability using MCMC. Another

important output of BEAST 2 is the demographic trend, represented by the posterior distribution of the effective population size N_e through time in the form of a Bayesian Skyline Plot (BSP). BEAST 2 was also employed to calculate Bayesian age estimates of tree nodes, by using as priors radiocarbon dates of ancient individuals, if available, and previously estimated MRCA ages of the major branches. BEAST runs were performed with complete mtDNA sequences under the HKY substitution model (gamma-distributed rates plus invariant sites) with a fixed molecular clock as in Brandini et al. [108], and setting as prior the clock rate considering the ones published in [55, 96]. The chain length was set to 10 million iterations with samples drawn every 1000 MCMC steps, after a discarded burn-in of 10%. The BEAST trees were summarized into a single "target" tree using TreeAnnotator and were visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). The BSPs were obtained with Tracer v1.7.1, but then converted into more realistic trends considering a generation time of 25 years [108].

6.1.3 Modern genome-wide data

A total of 135 individuals were genotyped with the Affymetrix Human Origin chip (~629K markers) in outsourcing at the Institute of Healthcare Research in Santiago de Compostela (CEGEN).

6.1.3.1 PCA

PCAs were performed using the "smartpca" program from the software package EIGENSOFT v7.2.0 [56]. Ancient data, presenting large amounts of missing data, were instead projected onto the modern variability, using the *lsqproject* and *autoshrink* options. Several PCAs were computed considering ancient and modern datasets and different sub-datasets. Outlier individuals were excluded from the downstream analyses.

6.1.3.2 ADMIXTURE clustering

Before running ADMIXTURE analyses, different datasets were pruned by removing SNPs in LD. This was done with PLINK v1.9 [242], applying the flag *--indep-pairwise 200 25 0.4*, with the values indicating the window size in kb, the base count to shift the window at the end of each step, and the pairwise r^2 threshold, respectively. For each step, pairs of SNPs in a window with r^2 higher than the threshold are noted, and the SNPs are pruned from the

window until no such pairs remain. After pruning, ADMIXTURE v.1.23 [58] was run performing ten independent runs for each K, from K1 to K20, and using the --cv flag to compute a 5-fold cross-validation error. This was used to select the best K, that is the one with the lowest error. The tool CLUMPAK [243] was used to combine the different runs and the software DISTRUCT [244] was employed to find the best alignment of the results.

6.2 Ancient DNA

In the works presented in this thesis, a total of 31 ancient DNA sequences have been produced. Among these, 19 ancient mitogenomes were obtained from 28 pre-Roman individuals from the necropolis of *Plestia*, located in East Umbria (central Italy). Four direct radiocarbon dates confirmed the age estimated from the archaeological context, placing the remains at the end of the 7th cal. century BCE. In addition, 12 low-coverage genomes were obtained starting from an original collection of 20 human remains (13 pre-colonial and seven colonial) retrieved from seven different archaeological excavations along the Pacific coast of Panama City and radiocarbon dated from 603 to 1,430 CE.

6.2.1 Low-coverage shotgun sequencing

The petrous part of the temporal bone and well-preserved molar teeth have been chosen as preferred anatomical element for DNA extraction, carried out with a silica-based method, which allows to recover very short DNA molecules, in dedicated clean rooms and following published protocols [214, 217, 245, 246]. Genomic DNA extracted from ancient remains was used to prepare DNA libraries, following a custom double-indexing protocol optimized for ancient DNA, and then shotgun-sequenced on Illumina platforms, as described in [5], and detailed in the Ph.D. thesis of Dr. Marco Rosario Capodiferro.

6.2.2 Ancient mtDNA analyses

After shotgun sequencing, mtDNA reads were extracted by mapping the raw data to rCRS [82], since the mtDNA sequence included in the GRCh37/ UCSC hg19 human reference genome is a Yoruba sequence, different from rCRS, commonly used in mtDNA phylogenetic studies. Since the release of the UCSC hg19 assembly, the mtDNA sequence has been replaced only in GenBank with rCRS. After quality control, trimming and adapter removal,

processed reads were aligned to rCRS using the BWA v0.7.17 aln/samse algorithm [230] and then realigned with CircularMapper [247], a specific tool to improve mapping on circular genomes. Duplicate reads were removed with Picard MarkDuplicates v2.21.6 (https://github.com/broadinstitute/picard) and the BAM files were further filtered and processed with SAMtools v1.9 [231], by keeping only reads with minimum mapping quality of 30 and positions with a minimum depth of 1.

Moreover, for some ancient individuals, mitochondrial DNA data was also produced using a capture technique (in collaboration with Prof. Caramelli's group at the University of Florence), which enriches DNA libraries for human mtDNA in a bead-capture method using LR PCR products as bait for hybridization [248, 249]. After enrichment, libraries were sequenced on an Illumina MiSeq platform using a paired-end approach. Raw paired-end reads derived from captured mitogenomes that overlapped for at least 11 bases were merged using the software ClipAndMerge v1.7.7 [247] and then filtered by quality trimming and adapter removal. Clean reads were filtered as above and final mtDNA BAM were merged with shotgun mtDNA BAMs using SAMtools *merge*. Two strategies allowed us to define the haplotypes. First, variant calling was performed using BCFtools v1.10.2 [231] and VCFs were filtered with VCFtools v0.1.16 [250]; haplotypes were then refined by manually checking the alignments using SAMtools *tview*. Second, after a first classification, if indel-related issues emerged (since some are diagnostic of specific lineages), cleaned reads of ancient individuals were realigned to modern mitogenomes belonging to the same haplogroup. The alignment was performed as above, using BWA aln/samse and then CircularMapper. Consensus sequences were generated using the same quality and coverage filters as before and then compared to rCRS to obtain the final haplotypes. As for the modern mtDNA haplotypes, haplogroup classification, based on PhyloTree build 17 [69], was achieved using HaploGrep 2 v2.1.21 [227].

6.2.3 Contamination estimates

As previously described (see chapter 4.6 Ancient DNA), after death DNA starts to become fragmented and damaged, e.g., due to deamination events occurring at the ends of DNA molecules. These features can complicate sequencing and data analyses, but they also are useful to effectively identify authentic aDNA reads. Modern DNA contamination is another issue that must be dealt with before conducting reliable analyses on aDNA data. Among the many available methods to estimate present-day contamination, three approaches were applied here, two based on mtDNA and one based on chromosome X. The rationale behind this choice is that all aDNA genomes were low coverage, preventing us to use autosome-based contamination estimates. In particular, the mtDNA is very useful to this purpose since it is

present in multiple copies in cells and much smaller in size than the nuclear genome, allowing to have enough coverage after sequencing to perform contamination analyses. For all three methods, a threshold of 3% of contamination was applied, considering as contaminated, and therefore removing from downstream analyses, aDNA sequences with estimates above this value.

The first tool used has been *schmutzi* [140], a set of programs that jointly estimate modern human contamination and reconstruct the endogenous mitochondrial sequence by considering both deamination patterns and fragment length distributions. It is worth mentioning that, since this tool is based on the detection of deamination patterns, aDNA data fully treated with uracil-DNA glycosylase to remove typical aDNA damage are likely to negatively impact these contamination estimates. *Schmutzi* can be applied to ancient data with both low and high levels of contamination. In details, *schmutzi* assumes that contaminant modern reads are not deaminated. Based on this, it measures the rates of deamination for the endogenous reads and compare them to the deamination rate of all fragments in the dataset, providing a rough contamination estimate. As an example, if the endogenous reads have 30% of deamination level, but the observed rate for all reads is 15%, the contamination estimate is 0.5. This first estimate is used as a prior in the next step, an iterative process, which includes the following three phases: (i) The endogenous consensus is called; (ii) the consensus called in the first step, together with measures of deamination rates and fragment length distribution of endogenous and contaminant fragments, are used to re-estimate contamination; (iii) the most likely contaminant from a non-redundant database of 197 human modern mitogenomes is estimated by determining the most likely contamination rate using sites where endogenous and contaminant genomes differ.

The second method used to estimate modern contamination on mitochondrial DNA data was *ContamMix* v1.0-10 [251]. This software package provides the maximum *a posteriori* probability of the consensus sequence being authentic. *ContamMix* assumes that the sequence coverage is high enough to call the true endogenous mtDNA consensus sequence, since contamination estimate is based on this reconstructed consensus. Moreover, it assumes that contamination in the data should not exceed 50%. The inferred fractions of exogenous mitochondrial sequences correspond to the amount of contamination. In detail, a consensus sequence is built from the alignment by running ANGSD v0.923 [252] with parameters *-doCounts 1* and *-doFasta 2*, therefore using a majority rule. Only reads with mapping quality higher than 30 and base quality greater than 20 are kept. In addition, only positions with a minimum depth of 5 are considered. Original reads that mapped uniquely to the mtDNA reference sequence are converted to FASTQ and then re-mapped to this consensus. The consensus sequence and a panel of 311 worldwide modern mtDNA genomes [253], serving as a source of potential

contaminants, are aligned using MAFFT v7.475 [254, 255]. Finally, *contamMix* is run using as inputs both this multiple sequence alignment as well as the new BAM file produced. It uses a MCMC framework to estimate the level of contamination. Five independent Markov chains are run from different random starting parameters, which are then used to test the Gelman-Rubin convergence [256]. Each chain is usually run for 50,000 iterations. Default values are used for other parameters. The results are checked by monitoring the Gelman diagnostic to confirm convergence. Generally, for mtDNA-based methods the contamination rate is acceptable until a threshold of 3% or 5% [5, 220].

The third tool was instead based on the X chromosome [257]. This method relies on the fact that males are hemizygous for X-linked *loci* outside the pseudo-autosomal regions, therefore making multiple alleles in these *loci* attributable to either errors or contamination. This contamination estimate can be used as a proxy for nuclear contamination in ancient male individuals. This software can estimate modern human contamination in low-coverage (as low as 0.2-0.5X on the X chromosome when contamination is below 15-25%) sequencing data. *ContaminationX* models base counts as a function of an error rate estimated from the data, the allele frequencies in the contaminant populations, and the contamination fraction, which is estimated by a maximum likelihood optimization. In details, ANGSD v0.923 [252] is first used to compute the allele counts on the X chromosome. This software is also applied to exclude pseudo-autosomal regions and then to estimate contamination fraction on reads with mapping quality greater than 30, base quality greater than 20, and minimum depth of 2X [221], using as a potential contaminant source the HapMap CEU allele frequencies [258]. The final estimation is carried out by *contaminationX* considering only individuals with a minimum number of 100 overlapping SNPs [221] to the potential contaminant reference panel. Contamination is estimated with a maximum likelihood approach, setting to 1,000 the maximum number of jackknife samples used for estimating standard errors and considering the estimates from the Two-consensus method. The threshold value for contamination estimated on the X chromosome is usually set at 3% [220].

7. Mitochondrial DNA analyses of modern and ancient populations

Most of the projects presented in this thesis employed mtDNA sequence data to reconstruct the genetic history of human populations in different geographic contexts, from Eurasia (chapter 7.1) to the Americas (chapter 7.2). Hereafter, the main conclusions of each work are presented together with a more detailed description of the main project, which is focused on Panama and describes the uniparental gene pool of admixed and Indigenous population groups currently living in the Isthmus.

7.1 Eurasia

In the works presented hereafter, a total of 2,965 control-region sequences and 338 complete mitogenomes from Eurasian populations have been analyzed.

7.1.1 The mitogenome portrait of Umbria in Central Italy as depicted by contemporary inhabitants and pre-Roman remains

The Italian Peninsula played a pivotal role in human migrations around the Mediterranean Sea, as testified by the higher degree of its current genomic variability compared with other European populations [259–264]. This complexity results from several different inputs that shaped the gene pool of this region since the Upper Paleolithic. Inferring the contributions of each process is further complicated by similar (or partially overlapping) migration events from, to and even within the Italian Peninsula, often separated by short time frames. It is generally agreed that the ancestral contribution is evident in the ancient Italic people known as Latins, while the invasions after the fall of the Roman Empire did not significantly alter the already established Italian gene pool [259, 263]. Several studies based on autosomal and uniparental markers tried to identify a clear genetic pattern able to discriminate northern, central, and southern Italian populations [263, 265–269]. Most of these studies were performed on a large geographic scale and mainly focusing on modern populations. So far, only a few microgeographic studies have been conducted in central Italy, specifically on Etruscans (in Tuscany) and on Picentes (in Marche) [270–277]. However, Umbria, another crucial region in central Italy, is still unexplored. The origins and genetic affinities of the ancient *Umbri*, traditionally considered an indigenous and very old population [278], are still debated. During the Early Iron Age (9^{th}-8^{th} centuries BCE), the *Umbri* were among the first communities with strong and well-defined cultural

identities in central Italy. They originally occupied today's eastern Umbria and then extended to the western part of the region, later occupied by the Etruscans. After 260 BCE, Umbria was already under the full control of Rome [279]. An important necropolis in east Umbria is in the so-called *Plestinam Paludem* (now Colfiorito, up in the Apennines). The *Plestia* plateaus represented an obligatory passage way in the trans-Apennine routes, but stable settlements have not been attested before the beginning of the Iron Age [280].

In this work [1], we produced and analyzed 198 complete mtDNAs from modern Umbrians (191 sequenced here for the first time), selected from a larger dataset of 545 control-region sequences, to investigate mtDNA variation in a microgeographic context and to obtain new insights of the maternal genetic history of Umbria. These data were integrated with 19 ancient mitogenomes from Iron Age *Umbri Plestini*, who were buried in the Colfiorito necropolis (Figure 14).

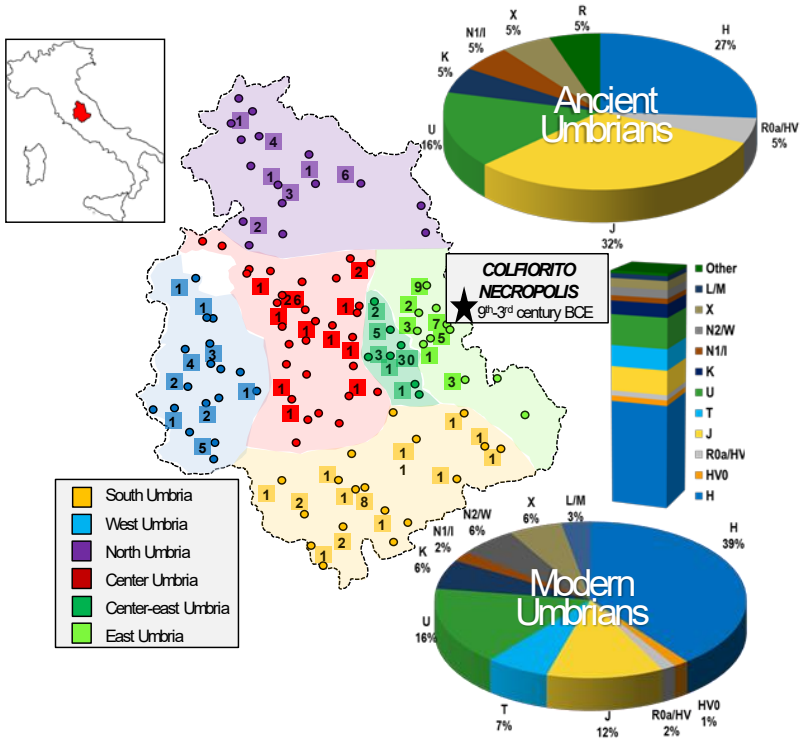


Figure 14. Geographic origin and Hg classification of ancient and modern Umbrians. The region is divided into six differently colored areas. Points represent the geographic origin of all modern individuals (N = 545), while complete mtDNAs are reported in squares (N = 198). Pie charts summarize haplogroup distributions considering complete mitogenomes of ancient (N = 19) and modern individuals, while the bar plot represents control-region data of the overall modern dataset. The location of the Colfiorito necropolis is indicated by a star. Adapted from Figure 1 in Modi et al. [1] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The 545 control-region sequences of modern Umbrians clustered into 369 haplotypes (haplotype diversity = 0.994), showing an extensive mtDNA variation in present-day Umbria. Most (97%) of the dataset was represented by western Eurasian lineages, with a rather homogenous distribution. A notable exception was Hg J, which has significantly higher frequency (30%) in eastern Umbria. PCA (Figure 15) comparing our dataset with a large Eurasian dataset (15,972 control-region seqeunces) [270], showed that east Umbria clusters together with eastern Europe, with major contributions to this clustering coming from haplogroups U4 and U5a (inset of Figure 15).

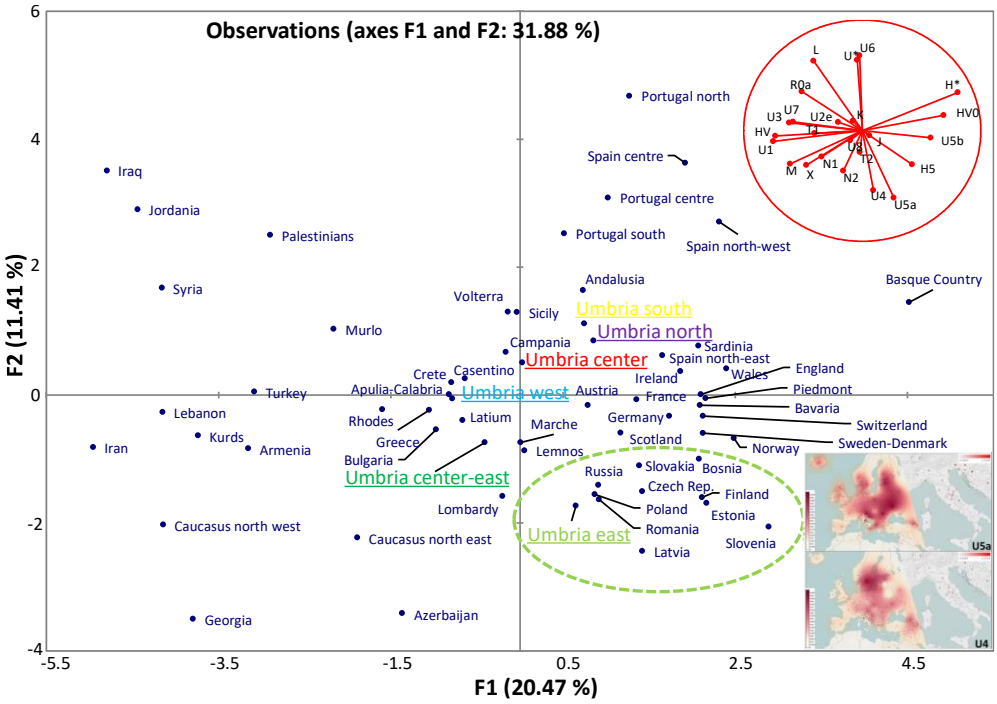


Figure 15. PCA of Eurasia based on haplogroup frequencies from control-region data. The inset shows the frequency distributions of U4 and U5a in western Eurasia (left side) as well as in Italy (right side). Adapted from Figure 3 in Modi et al [1] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

A total of 198 complete mtDNAs were sequenced randomly selecting representative individuals from the six established regional divisions (see Figure 14) and considering current population densities. By using the capture technique, we recovered mtDNAs from 19 ancient individuals (dated from the early 9th to the late 3rd century BCE) from the necropolis of *Plestia* (Figures 14 and 16), with an average depth of coverage ranging from 5.86X to 50.98X. By comparing the mitogenomes of modern Umbrians and pre-Roman individuals, we found a continuity in the eastern part of the region, testified by a similar high frequency (~30%) of haplogroup J in both modern and ancient sequences, and probably maintained by geographic isolation. However, such

a continuity is probably extended over the entire region, since almost all present-day lineages were already present in pre-Roman times and more than half of the ancient individuals share terminal branches (H1e1, J1c3, J2b1, U2e2a, U8b1b1, K1a4a) with modern Umbrians, all dated back to the Holocene (Figure 16).

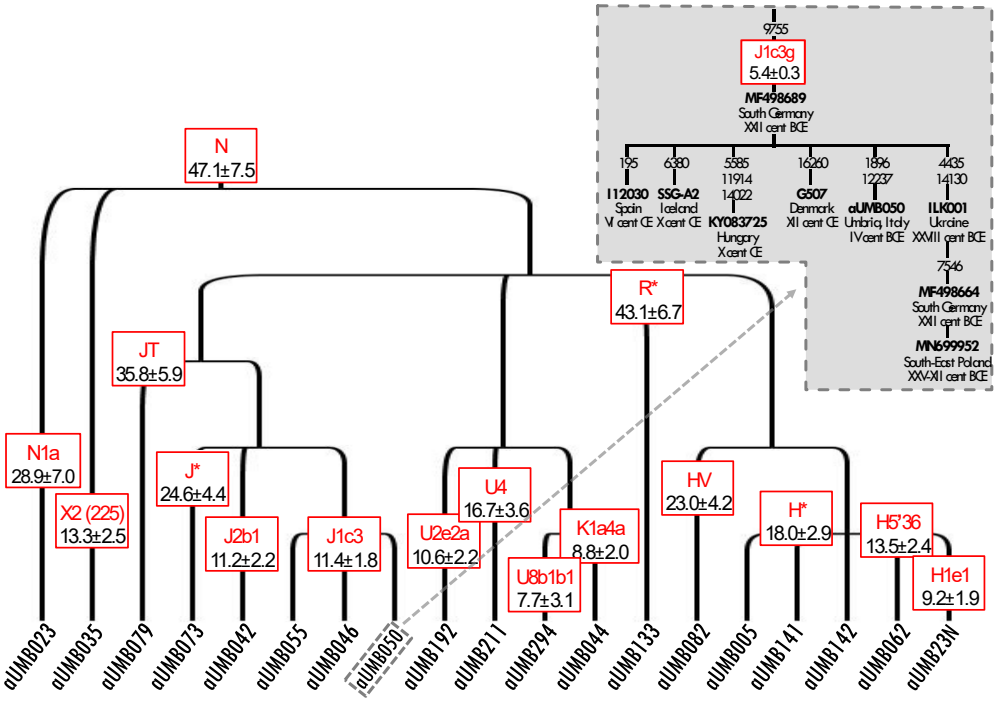


Figure 16. Schematic phylogeny of ancient Umbrians. Bayesian ages refer to the MRCA shared with modern Umbrians. The inset highlights the closeness of the Umbrian J1c3g mtDNA to other available ancient mitogenomes from western Eurasia. Adapted from figure 5 in Modi et al. [1] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

These specific lineages were also found in published ancient individuals from different regions of western Eurasia and northern Africa and covering a long-time span. J1c3g could be considered a paradigmatic example of these heterogeneous connections, as attested by its aDNA tree, which includes the ancient Umbrian aUMB050 and other eight published ancient mitogenomes [281–287] (inset of Figure 16). These include Neolithic, eastern Europe Bronze Age, Bell Beaker, and more recent Medieval individuals. These results suggest that specific mtDNA variants have been brought into the region by the ancestors of ancient *Umbri*, coming from various sources at different times. These connections range from Neolithic farmers, spreading along the Mediterranean, to Bronze Age and Medieval connections with central-eastern Europeans, possibly including nomadic groups (Yamnaya) from the Pontic-Caspian steppes.

This microgeographic and diachronic mtDNA portrait of Umbria fits well with recent Y-chromosome and genome-wide data on the entire peninsula [263, 266, 288] and agrees with historical sources that list the *Umbri* among the most ancient Italic populations, genetically and linguistically distinct from the neighboring Etruscans.

7.1.2 Mitochondrial DNA footprints from Western Eurasia in modern Mongolia

Archaeological evidence [289, 290] and genetic studies on modern and ancient populations provided information on the complex Mongolian past [291–300]. These studies revealed at least four ancestral sources that arose in Mongolia through the Neolithic. Two were associated to hunter-gatherers from northeast Asia and northern Eurasia, one was associated with the eastward expansion of the Yamnaya culture (Afanasievo culture), and the fourth was constituted by a mixture of Yamnaya culture and European farmers [301, 302]. During Middle and Late Bronze age (~1900-900 BCE) dairying and nomadic herding was widespread. Development of large-scale polities in this period led then to dynasties and empires in the Early and Late Medieval times. Since then, the entire Eurasian Steppe became an important crossroads through the Silk Road, which played a major role in the economic, demographic, and cultural processes shaping the history of several Eurasian populations [293]. Another important event in the history of Mongolia was the rise of the Mongol empire in the late 12th century CE with the ascent of Genghis Khan (“Universal Ruler”). At its peak (1206–1368 CE), the empire stretched from present-day Poland in the west to Korea in the east, and from Siberia in the north to the Gulf of Oman and Vietnam in the south. During this time, the so-called *Pax Mongolica* allowed a period of commercial, cultural, religious, and scientific exchanges between western and eastern populations, including trades between nomadic groups and urban centers [303].

MtDNA studies contributed several pieces of information to disentangling the genetic history of Mongolia. Evidence deriving from the mtDNA Hgs shared between Afanasievo and Yamnaya people supports an eastward migration from the Pontic-Caspian steppes [186, 295]. The presence of a U5a1 mitochondrial haplotype in an Eneolithic grave, dated at ~3000 BCE and associated with the Afanasievo archaeological culture in the Khangai Mountains, attested the presence of people with "western" origin in the east of the Altai Mountains before the Bronze Age [304]. To further investigate the impact and legacy of eastern and western mitochondrial lineages on the gene pool of modern Mongolian populations, we analyzed the mtDNA profiles of 2,420 individuals from 20 different Mongolian provinces. In addition, the entire mitogenomes of 147 individuals, representative of different Hgs, were sequenced.

We assembled a control-region dataset of 2,420 sequences, 2,335 of which encompassing the complete HVSI region (and used for diversity indexes analyses), which revealed a high mtDNA variation heterogeneously distributed across the country (Figure 17A). The increase near the capital, Ulaanbaatar, could be probably explained by very recent migrations that may have hidden the original mtDNA pool. Other populations possibly remained more isolated due to geographic barriers, which reduced the number of different mitogenomes since ancient times (e.g., in the Khangai Mountains) or only recently (e.g., in the Khovd region). Most of mtDNAs belong to Hgs typical of eastern Asian populations whose frequency decreases from eastern to western regions (Figure 17B). An opposite pattern could be observed for western Eurasian lineages. Overall, both contributed to create the mtDNA differentiation currently detectable in Mongolia, as highlighted by the PCA (Figure 17C).

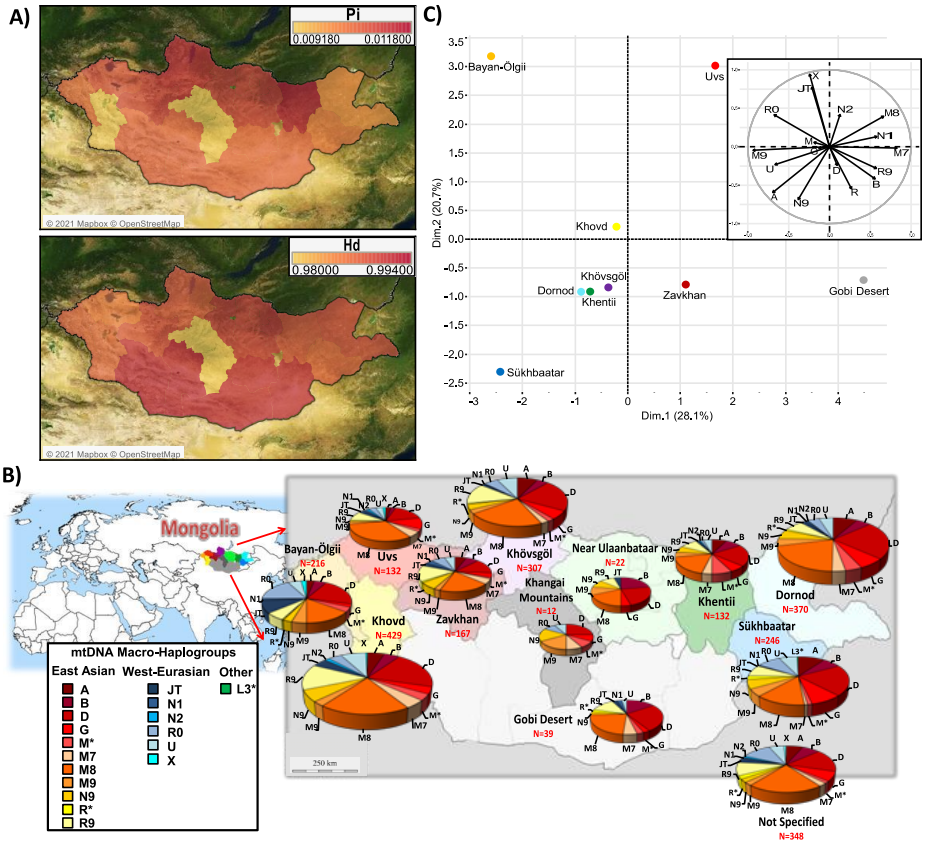


Figure 17. The mtDNA variation within Mongolia based on 2,420 modern control-region sequences. A) Map of genetic variability in each macro-area expressed as nucleotide diversity (Pi) and haplotype diversity (Hd). B) Pie charts showing the macro-Hgs distribution. C) PCA plot representing the genetic landscape of Mongolia based on macro-Hgs frequencies. The following groups were excluded: "Khangai Mountains", "Near Ulaanbaatar" and "Not Specified". Adapted from Figure 1 in Cardinali et al. [2] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The western Eurasian lineages JT, R0 (mostly made of H mtDNAs), and X determine the genetic distinction of the three westernmost provinces along PC2, with Bayan-Ölgii (encompassing Altai mountains) having an outlier position. The Altai mountains have traditionally been considered a genetic barrier to gene flows from the west until the discovery of ancient individuals with western haplogroups east of these mountains before Bronze age challenged this view [304]. These results confirm that a western influence can also be detected in modern mitogenomes. Different eastern Asian lineages characterize the southern regions, while the northeastern provinces (Dornod, Khentii, Khövsgöl and Sükhbaatar) cluster together, separately from the others, and are characterized by a high number of different mitogenome variants that arrived mostly from the surrounding eastern Asian countries. When evaluating the Mongolian gene pool in a Eurasian context, we observed a greater proximity to the surrounding East Asian populations, mostly driven by eastern Asian haplogroups. The legacy of western Eurasian lineages was less marked and more widespread. To deepen the understanding of mtDNA peculiarities of Mongolians, we sequenced 147 complete mtDNAs, 26 representative of eastern Asian lineages and 121 belonging to western Eurasian Hgs. In the phylogenetic tree of western Eurasian clades, many lineages coalesced during and soon after the last glacial maximum (LGM, ~25-15 kya; purple nodes) (Figure 18A-B). The BSP describes a demographic trend with two major increases of the effective population size (Figure 18C). The first one reflects post-LGM migrations from glacial refuges in western Eurasia, as also attested by haplotype sharing with Europe and Balkan populations, and by the high frequency of haplogroup H1, a genetic marker of the post-LGM expansions [305]. The second took place during the early Holocene and was probably facilitated by ecological changes associated with the Holocene climatic optimum (~10-6 kya) [306, 307], that was accompanied by the development of farming, pastoralism and more sedentary communities. A mixed ancestry between Yamnaya and European farmers was recently identified in Bronze Age Mongolians [301, 302]. We did not find western Eurasian sub-lineages specific to Mongolia. Therefore, most of the lineages detected in modern Mongolians actually evolved in western Eurasia. Some lineages started to coalesce in the early Bronze Age (~5 kya), while very few Hgs originated in more recent times (<3 kya). The lack of Mongolia-specific sub-branches might also suggest that the western Eurasian lineages arrived in the Eastern Steppe in more recent times. The ages of some western Eurasian lineages dates between 5 and 3 kya could be linked to Bronze Age migrations across the Eurasian steppes that probably involved also the Afanasievo first (~3300–2500 BCE) and later the Sintashta culture (~2100–1800 BCE).

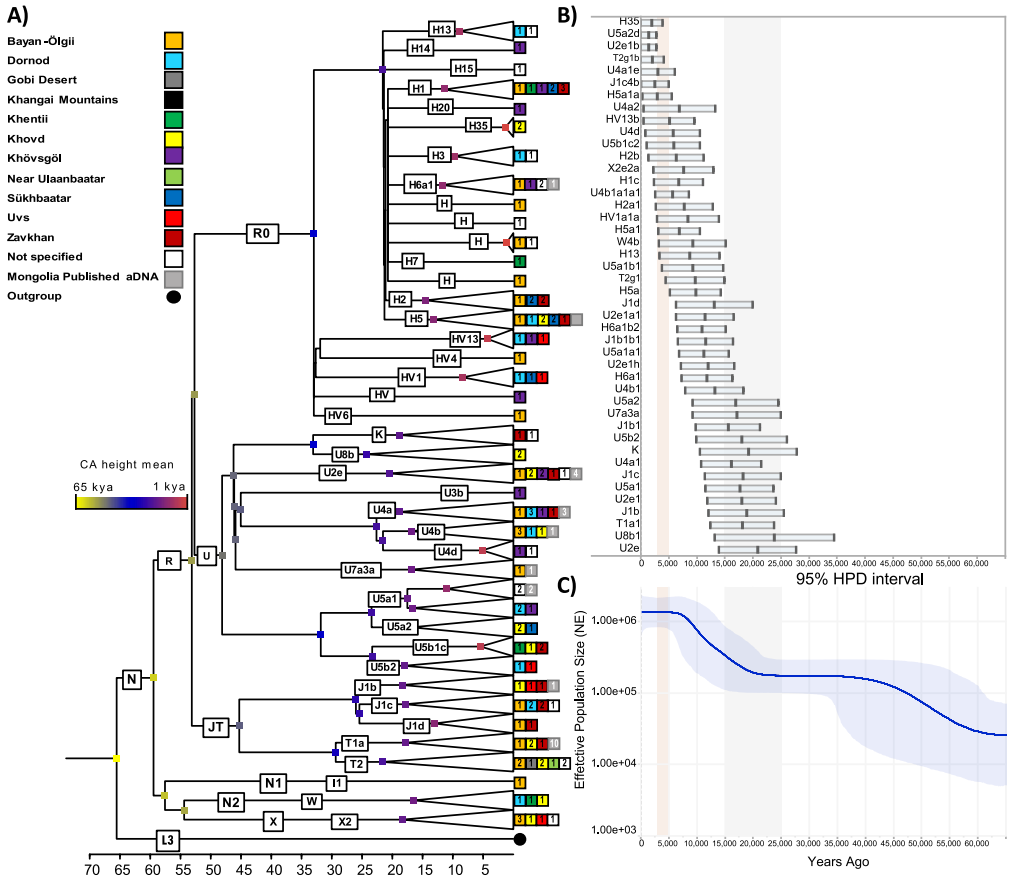


Figure 18. The mtDNA variation in Mongolia based on 121 (new) modern and 25 (published) ancient complete mitogenomes belonging to western Eurasian haplogroups. A) Bayesian tree with internal nodes colored according to common ancestor (CA) average age estimates. B) Age estimates and 95% high posterior densities. LGM timeframe is highlighted in grey, while the red shade indicates Bronze Age. C) BSP displaying changes in the effective population size through time considering a generation time of 25 years. Adapted from Figure 3 in Cardinali et al. [2] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Finally, by searching the available database of ancient mitogenomes for western Eurasian lineages identified in our modern Mongolians, we found 13 different sub-haplogroups in ancient Mongolians dated after the Bronze Age. They might testify for small population movements from the west less than 3 kya that can be probably related to commercial routes. The migration path from western Eurasia to Mongolia marked by some of these mitochondrial sub-lineages (H5a1, J1b2, T2g, U2e1b, U4b1a1a1, and U4b1a4) occurred ~2.5 kya, thus temporally and geographically overlapping with the Silk Route, while other sub-haplogroups, such as J1b1b1 and U2e1a1, seem to have arrived in Mongolia later.

In conclusion, we observed a clear genetic differentiation in Mongolia, which reflects a history of connections with eastern Asia as well as with western Eurasia. The prevalent Asian contribution reveals continuous connections with neighboring populations until recent times, likely facilitated by the so-called Genghis Khan's *Pax Mongolica*. As for the western Eurasian haplogroups, the analysis of complete mtDNAs highlights two major changes in the effective (female) population size. The less recent one started in the Late Pleistocene, before increasing in the early Holocene, and for the first time points to a mtDNA connection with post-glacial repopulation events involving western Eurasian refuges. Finally, a diachronic comparison with ancient mtDNAs links six mitochondrial lineages of present-day and ancient Mongolians with the timeframe and geographic path of the Silk Route.

7.2 The Americas

In the following mtDNA studies on American populations we produced a total of 2,452 control-region sequences. Moreover, in the works on Panamanian populations we also generated Y-chromosome data for 248 male individuals, as well as 84 modern and nine ancient complete mitogenomes (see chapter 8.1).

7.2.1 The mitochondrial DNA landscape of modern Mexico

After the first peopling [308–311] and the development of (maize) agriculture [309, 312–315] in South-Central regions, Mexico was characterized by population movements and growth of complex socio-economic, political, and religious civilizations [309, 315, 316]. Population decline and bottlenecks occurred before, and coincided with, European contact after 1519 CE [316]. Even if Indigenous populations recovered since the 17th century and remained largely isolated until the mid-18th century [312], the demography of Mexico was drastically reshaped following colonization. African slaves were brought in as a workforce and the pre-Columbian civilizations were eventually displaced and annihilated [317, 318]. Mixed populations increased during the 17th and 18th centuries [319]. Today, Mexico shows one of the richest ethnic and linguistic diversities of the world [311, 320]. Most of the Mexican population is largely formed by descendants of Indigenous Americans, Spanish (European) immigrants, and African slaves [321]. Considering that census data on ethnicity are not officially collected, the proportions and definitions of population groups are not consistent. Most people (~85–95%) identify themselves as a mixture of Indigenous American and European ancestry, while present-day Indigenous groups are defined by culture, traditions, language, and oral history. Recent estimates found 2% of the

population seeing themselves of African ancestry. For these reasons, Mexico has been, and remains, a preferential target for genetic studies, particularly on mtDNA. However, previous studies were either based on very short sequence fragments [318, 319] or on a small number of subjects [317, 322] from geographically restricted regions [323], or did not report the actual haplotypes [324]. To address these limitations, this study presents the first comprehensive overview of modern mtDNAs from Mexico reporting control-region data of 2,021 individuals from the general population across the country (Figure 19).

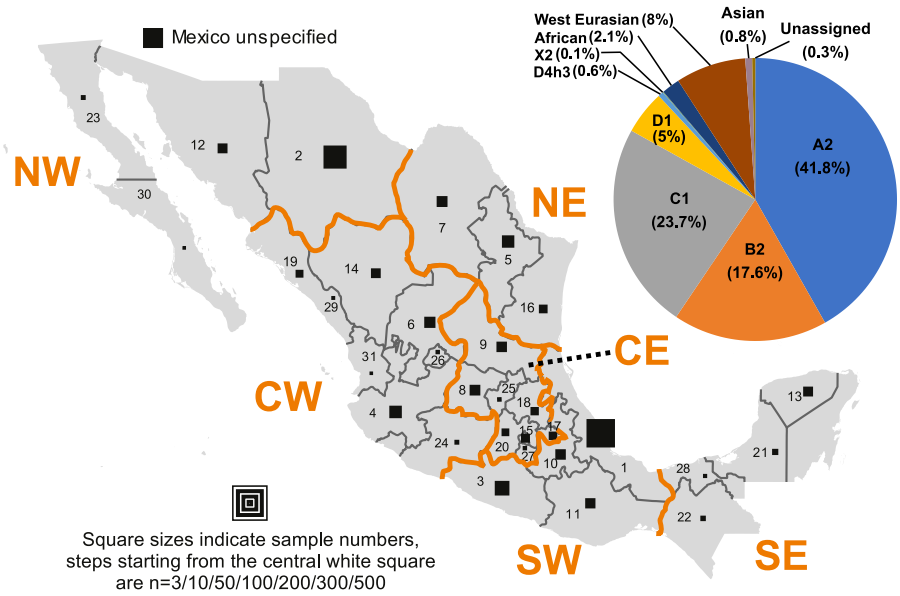


Figure 19. Geographic origin and haplogroup classification of the 2,021 Mexican individuals. The map shows the administrative units of Mexico and the individuals from each of the units analyzed in this study. The geographic subsets indicated by orange lines and text are Northwest (NW), Northeast (NE), Center-West (CW), Center-East (CE), Southwest (SW), and South-East (SE). The pie chart shows the proportions of haplogroups in the Mexican mtDNA pool. Figure 1 in Bodner et al. [3] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

All 2,021 haplotypes passed quality control on EMPOP, the forensic online mtDNA population database [325]. The final dataset contained 799 different haplotypes, 502 of which (62.8%) were unique (haplotype diversity = 99.6%). This dataset revealed a wide-ranging spectrum of mtDNA haplogroups present in the general country-wide Mexican population. The predominant proportions of clades were attributable to Indigenous American lineages and altogether comprised 1,804 (89.2%) of the individuals. All major pan-American haplogroups (A2, B2, C1b, C1c, C1d, and D1) [110] were highly represented (Figure 19). The high number of sub-lineages likely reflects the history of settlement, combining multiple populations with different origins, history, and bottlenecks [311]. Haplogroup A2 was the most prevalent with

845 individuals (41.8%), subdivided into 30 different sub-haplogroups. B2 was represented by 356 individuals (17.6%) and 21 sub-lineages. Hg C1 comprised 478 individuals (23.7%) and 19 clades. Lastly, 111 individuals (5.5%) fell into the D1 lineage with six sub-haplogroups. Moreover, two rare founder lineages, D4h3a and X2a [109], were found, represented by 12 (0.6%) and two (0.1%) individuals, respectively. A total of 162 individuals (8.0%) belonged to West Eurasian haplogroups (H, HV, J, K, R0, R6, T, and U), while 42 individuals (2.1%) were classified into sub-Saharan African lineages (L0, L1, L2, and L3). Finally, South, Southeast, and East Asian haplogroups were represented by six individuals bearing F1a1'4 and M lineages. The frequency of Indigenous lineages is in line with previous estimates in the general Mexican population [318, 319, 326, 327], although on a finer geographic scale some heterogeneity of frequencies was detected. The spatial distribution of haplogroups within Mexico was highly heterogeneous (Figure 20).

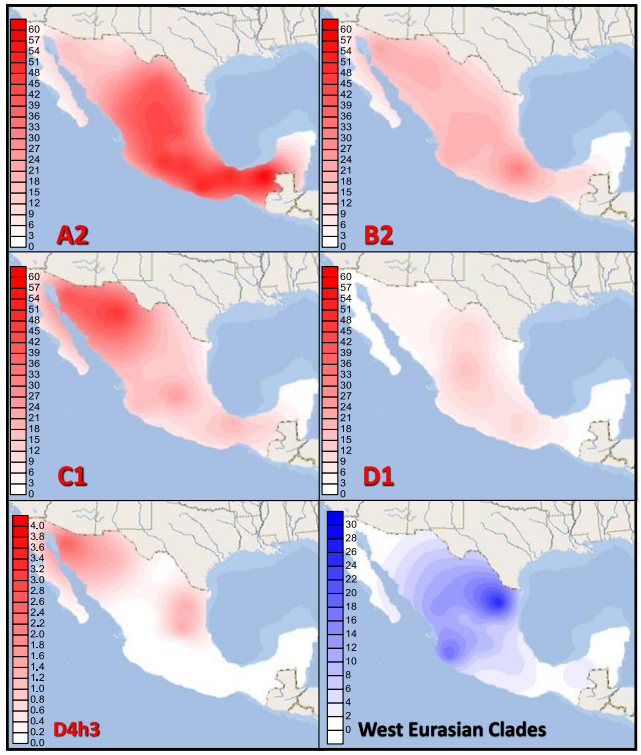


Figure 20. Spatial frequency (%) distributions of Indigenous American haplogroups A2, B2, C1, D1, and D4h3 and the combined West Eurasian lineages in Mexico. Adapted from Figure 2 in Bodner et al. [3] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The distribution of Indigenous lineages showed specific patterns along the country. Haplogroups A2 and B2 show a decrease of frequency from south to north, while C1 has an opposite trend, from northwest toward south. These

findings illustrate the peculiarity of lineage dispersal at a microgeographic level, since it contrasts with the overall gradients in the double continent [319]. D4h3a and X2a behavior was in line with previous reports on their specific distribution in the Americas [109]. African lineages were uniformly distributed, while the Western Eurasian Hgs were mainly found in the northeast. In general, this study shows overwhelming Indigenous American maternal legacy in the extant admixed Mexican population, with almost 90% of mtDNAs belonging to Indigenous lineages, in contrast with estimates from the Y chromosome [328]. This is explainable by the mode of European conquest that included “directional mating” of immigrant men with indigenous women, causing asymmetric genetic admixture in Indigenous and urban populations in and beyond Mexico [321, 329, 330]. Moreover, the haplogroup frequency patterns, corroborated by the haplogroup frequency based PCA revealing clear haplogroup structure, suggested the necessity of evaluating regional subsets to yield accurate dispersal patterns and forensic rarity estimates. Thus, database subsets containing northern (n = 545), central (n = 589), and southern (n = 679) haplotypes were compared. The relevance of regional databases was clearly proven by the fundamental differences revealed in the dispersal of haplotypes and singletons, which are of high relevance for forensic statistics [331]. Around 80.0% of each subset’s haplotypes were not shared at all among them. In total, 85.0% of the unique haplotypes per region were restricted to their region.

In conclusion, this study provides for the first time a comprehensive overview on the mtDNA variation in the present-day general population of Mexico using a large dataset covering the entire country. The outcomes confirm a slight genetic impact of European conquest in terms of maternal introgression, resulting in the preservation of the pre-Columbian pattern of mtDNA variation in Mexico, as confirmed by the heterogeneous spatial distribution of some lineages. Then, a microgeographic analysis argued in favor of regional databases at least for forensic genetic investigations. Finally, this study has also provided insights of a possible sex-based differential mobility and mixture that impacted cultural as well as biological survival in Mexico.

7.2.2 Weaving mitochondrial DNA and Y-chromosome variation in the Panamanian genetic canvas

The Isthmus of Panama was an obligatory passage for the first peopling of the Americas and played a pivotal role during the European colonization and the African slave trade [332–334]. After the first peopling of the Isthmus, attested ~16 kya [335, 336], and the establishment of agriculture from 8.6-5 kya [337, 338], three cultural regions could be distinguished in the Isthmus starting from ~3 kya. The Greater Chiriquí (corresponding to present-day western provinces and Indigenous *comarcas* of Panama, and eastern Costa

Rica) was the most coherent historical unit [339], speaking Nuclear Chibchan languages [340]. Greater Coclé (comprising present-day central provinces) was a culturally coherent unit, although it is not known whether it was linguistically united. Finally, the Greater Darién (corresponding to eastern provinces and Indigenous *comarcas*) was less coherent and was inhabited by people speaking Cueva, since 1500 CE. This vernacular was more likely a *lingua franca* for social and trading communication [341]. This cultural heterogeneity persisted upon the arrival of the Spaniards at the beginning of the 16th century CE [332]. European colonization had a severe impact on autochthonous populations, which experienced a decline due to infectious diseases [342], warfare and labor in mines and pearl fisheries [343]. Panama became a principal redistribution point for the slave trade and mixture among African and Indigenous peoples became common [344]. Consequently, according to the latest Panamanian census (https://inec.gob.pa/), the percentage of citizens who identify themselves as Afro-descendants is 9.2%, while 12.3% consider themselves Indigenous.

Recent genomic analyses of populations currently living in the Americas confirmed the existence of sex biases in the convergence of diverse ethnic groups during and after European contact [345–347]. This was often documented by the differential inheritance of uniparental lineages [348–351]. In Panama, previous studies have shown greater Indigenous maternal legacy [352] in the general population and a much lower Indigenous paternal component [353], which is mostly characterized by Western Eurasian Y-chromosome lineages. This sex bias has been interpreted as the result of asymmetric coupling between European males and Indigenous American as well as African women. A related hypothesis suggests that more Indigenous men than women perished or were deprived of reproductive rights after contact, due to warfare and forced displacements of enslaved Indigenous males. Another peculiar feature of these previous studies concerned the general population of Panama presenting a lower African male than female component, which contrasts with historical records of approximately two men for every woman being involved in the trans-Atlantic slave trade [354, 355]. However, these previous studies did not consider any self-reported ethnic affiliation. Here, we generated mtDNA control-region data for 431 maternally unrelated individuals (301 males and 130 females), which were sampled from the general population (N = 210), five Panamanian Indigenous groups (N = 200; Naso, Bribri, Ngäbe, Guna, and Emberá), and two admixed groups, Mestizo (N = 10) and Moreno (N = 11). Y-chromosome haplogroups were classified for a subset of 248 Panamanian unrelated males, belonging to the same population groups.

The classification into mitochondrial haplogroups (Figure 21) revealed a prevalence of the Indigenous pan-American founding lineages (A2, B2, C1, D1), totaling 86.3% of the entire dataset and 75.7% of the general population, which is in agreement with Perego et al. [352]. All individuals from Indigenous

American (IAm) populations belonged to Indigenous lineages, except for one Emberá individual whose mtDNA was identified as haplogroup R*. A2 is the most represented IAm haplogroup (51.7%), followed by B2 (27.1%), C1 (6%), and D1 (1.4%; found only in the Emberá group). A2 could be further classified into two main sub-lineages: A2w represents 11.8% of the entire dataset and is found among the Ngäbe (43.7%), the Mestizo (60.0%), and the general population (6.7%); A2af1 (24.6% of the total) is found across all Indigenous groups (except for the Emberá), although with significant differences (Chi-square test, *p-value* < 0.01), and reach the highest frequencies in Guna (52.0%) and Naso (56.3%) populations. This lineage is found in one Mestizo individual and is also the most represented in the general population (26.2%). Haplogroup B2 is found in all groups, with significant differences among the IAm populations (Chi-square test, *p-value* < 0.01), and the highest frequencies in Emberá (46.9%) and Bribri (64.3%). C1 and D1 are found in the eastern populations (Guna and Emberá), with D1 present only in the latter and C1 equally distributed (Chi-square test, *p-value* = 0.83) in both groups (~20–22% in each population).

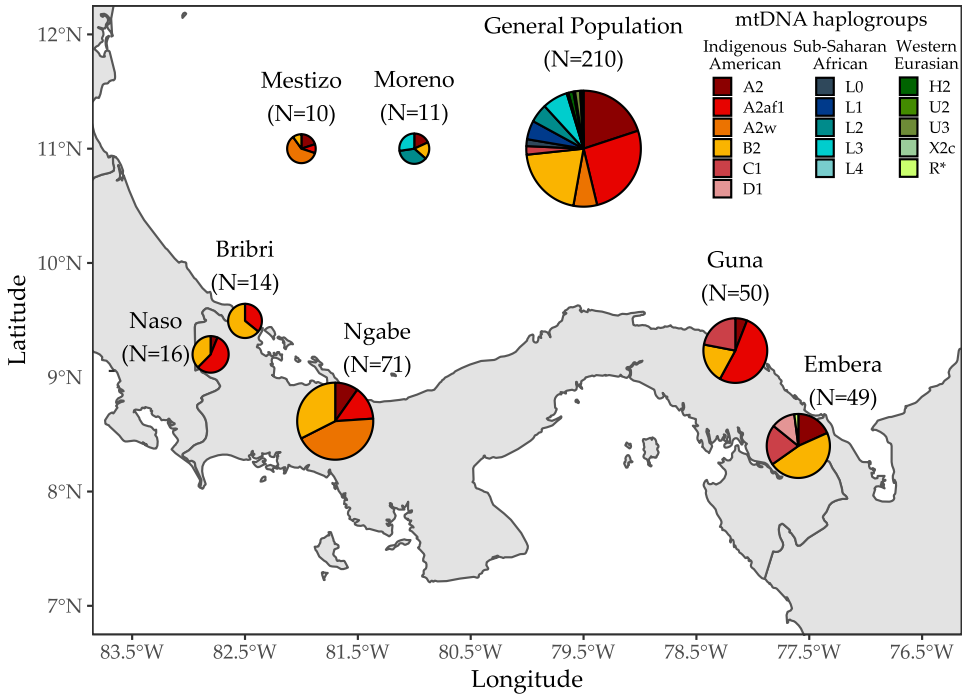


Figure 21. Map showing the Hg distribution of the 431 mtDNAs according to the self-reported affiliation of the study participants. Location of Indigenous groups corresponds to their specific Indigenous *Comarca* (Naso Tjër Di, Ngäbe, Guna, and Emberá). As for the Bribri, most of them live within Costa Rican borders, and about 3,000 are settled in Panama. The size of each pie chart is proportional to the number of individuals from each group. Adapted from Figure 1 in Rambaldi Migliore et al. [4] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Non-Indigenous lineages (13.7%) are represented by Western Eurasian (WEu; 2.3%) and sub-Saharan African (SAf; 11.4%) Hgs. WEu lineages are only found in the general population (4.3%), except for the previously cited R* mtDNA from Emberá. The Mestizo do not present any WEu clade. Conversely, SAf haplogroups are found in the general population (20.0%) and in the Moreno group (63.7%). In the latter group, SAf maternal ancestries are represented by L2 (36.4%) and L3 (27.3%) lineages.

Y-chromosome haplogroups were classified for a total of 248 Panamanian unrelated males (Figure 22). The IAm component is represented by haplogroups Q-M848 and Q-Z780. For Q-M848, only the Q-M925 branch is observed, with ~82% of individuals belonging to its sub-branch Q-Y12421. The remaining Q-M848 Y chromosomes are negative for downstream markers and therefore are classified as Q-M848*. As for Q-Z780, ~50% of Y chromosomes belong to the Q-SA02 sub-branch. The remaining Q-Z780 chromosomes, negative for downstream markers, are reported as Q-Z780*. The general population is mostly represented (69.8%) by WEu haplogroups, the most frequent being the Iberian R1b-S116 (28.2%). The second most frequent source in the general population is represented by SAf haplogroups (16.8%), with E-M2 being second in frequency (15.4%), after R1b-S116. Lastly, IAm Q sub-lineages accounts for 13.4% and are mainly represented by Q-Y12421 (6.7%). Among the admixed groups, Mestizo accounts for, in order of frequency, IAm, WEu, and SAf contributions, while among the four Moreno Y chromosomes, three were E-M2 and one was J2-M172, thus identifying a predominantly SAf paternal ancestry (E-M2). The Indigenous populations have mostly retained an IAm paternal ancestry (80.9%). Only the Bribri differ from this trend, represented only by the IAm Q-M848* (25%). The most frequent sub-haplogroup among IAm groups is Q-Y12421, which reaches its highest frequencies (~31%) in the Guna and the Naso. Q-SA02 is only observed in western groups (Naso and Ngäbe), while all the Q-Z780 in the Emberá remain classified as Q-Z780*. No Q-Z780 Y chromosomes were identified in the Guna. Q-M925* is almost exclusively found in the Ngäbe, except for one Guna individual.

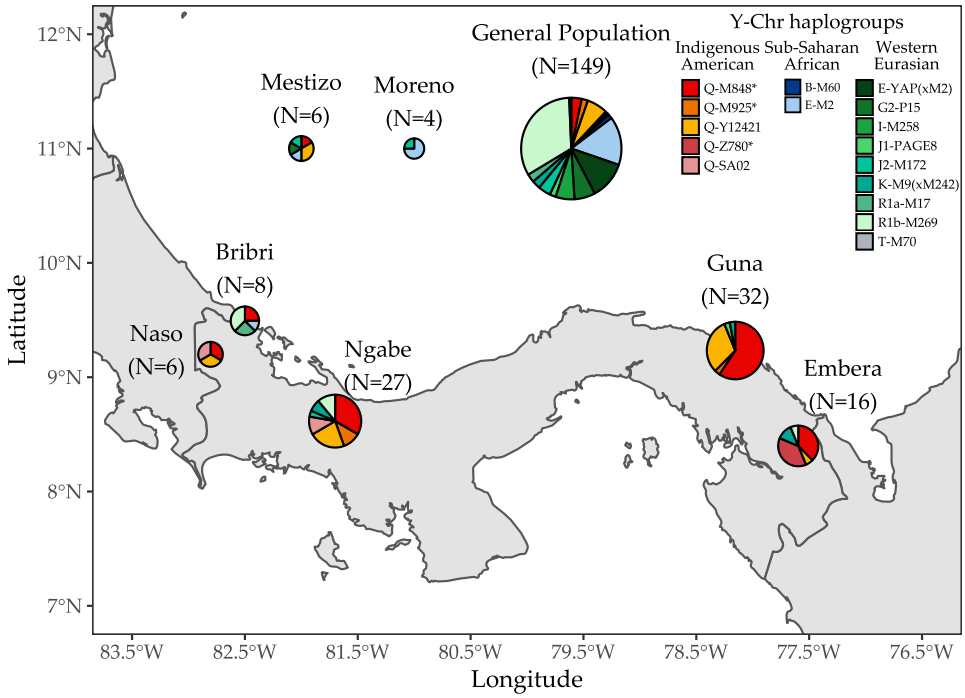


Figure 22. Map showing the haplogroup distribution of the 248 Y chromosomes according to the self-reported affiliation of the study participants. See Figure 21 for further details. Adapted from Figure 2 in Rambaldi Migliore et al. [4] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Both uniparental systems showed significant differences in haplogroup distributions among all Panamanian Indigenous groups (Chi-square test, *p-values* < 0.001). A paradigmatic example is the mtDNA haplogroup A2w, which is only present in the Ngäbe group and previously found in North and Central American modern individuals [5, 37, 356–360] and, more recently, in an ancient pre-Hispanic individual excavated in Panama City [5]. Likewise, the male paragroup Q-M925* is exclusively found (except for one Guna individual) in the Ngäbe. The mtDNA haplogroups C1 and D1 are found in Emberá and Guna (only C1) in the east and not observed in western groups. Similarly, Q-Z780 Y chromosomes from the west (Naso and Ngäbe) all belong to the sub-lineage Q-SA02, while those in the east (Emberá) remain classified as Q-Z780*.

To compare Y-chromosome and mtDNA data (Figure 23), we restricted the total mtDNA dataset of 431 individuals to the one used for Y-chromosome analyses (248 males). We did not find any statistically significant difference when comparing the mtDNA haplogroup distribution between the two datasets (Fisher's exact test, *p-value* = 0.998).

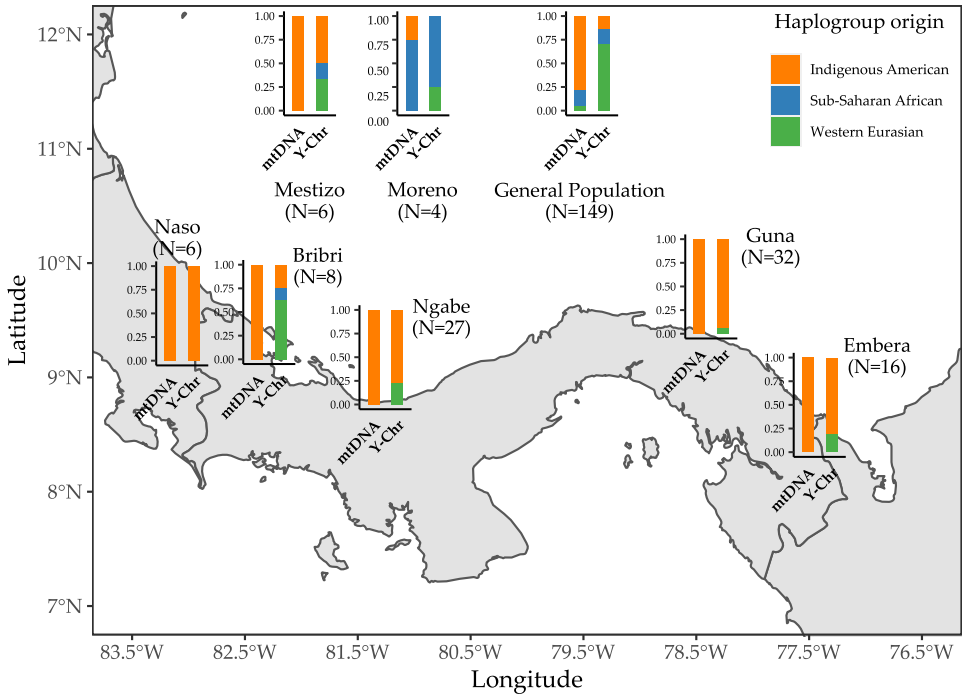


Figure 23. Map showing the distributions of mtDNA and Y-chromosome haplogroup origins (indicated by different colors) among 248 Panamanian males. See Figure 21 for further details. Adapted from Figure 3 in Rambaldi Migliore et al. [4], under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

IAm groups reach 100% frequency of IAm mtDNA lineages, but show an important WEu Y-chromosome contribution, which is predominant in the Bribri. The Moreno show the lowest proportion of IAm lineages for both systems, being SAf haplogroups predominant. The Mestizo and the general population are characterized by similar patterns, with predominant IAm mtDNA lineages and substantial WEu and SAf Y chromosomes. Here, in contrast to previous estimates [352, 353], the general population shows the same proportion of SAf lineages (16.8%) for both systems.

These analyses confirmed a sex bias in the general population currently living in Panama, with high frequencies of IAm mtDNAs (75.7%) and much lower IAm Y chromosomes (13.4%). This trend confirms patterns observed in other population contexts around the Isthmus, such as Mexico [3, 312, 319, 328] and Colombia [348]. However, this is not proved for other admixed populations with lower frequencies of IAm mtDNA lineages [361]. These patterns can be explained by statistically inadequate datasets and low-resolution uniparental screening of only modern individuals, but they surely testify to the complexity of admixture processes between populations with various cultural and biological backgrounds. An additional proof of this complex scenario is the different patterns detected in the IAm groups

currently living in Panama. The Indigenous mtDNA lineages frequency reaches 100%, while the Y-chromosome haplogroups range from 25% in the Bribri to 100% in the Naso. Similar patterns have been detected in IAm populations currently living both to the north (Costa Rica, Nicaragua, and Mexico) [312, 328, 346] and to the south (Colombia) [350] of Panama, as well as in other regions from all over the double continent [362]. Within the Isthmus, we pointed out a different post-contact impact of allochthonous Y-chromosome lineages on the Indigenous genomic pool of the western Panamanian groups that were the most homogeneous in pre-Hispanic times [5]. This may be due to a sample bias, particularly for the groups with less than ten individuals in the Y-chromosome dataset (Bribri and Naso), but cultural implications should also be considered. It is possible that Bribri women intermarried or otherwise coupled with men of both WEu and SAf origins, and the newborns were considered members regardless of their paternal origin. We also re-evaluated estimates of SAf contributions to the general population when comparing the mtDNA and Y chromosome. The same values for both uniparental systems (16.8%) appear somewhat closer to historical data. These estimates are probably more accurate than those obtained in previous studies [352, 353] where the self-declared ethnic affiliation was not recorded during sampling. Therefore, it is likely that some individuals of the general population were actually members of the Indigenous communities, where the incidence of sub-Saharan Y-chromosome lineages is much lower (~1%).

We computed genetic diversity either as haplogroup frequency-based heterogeneity [235], or as intra-population pairwise genetic distance based on variable SNPs for both systems (Figure 24).

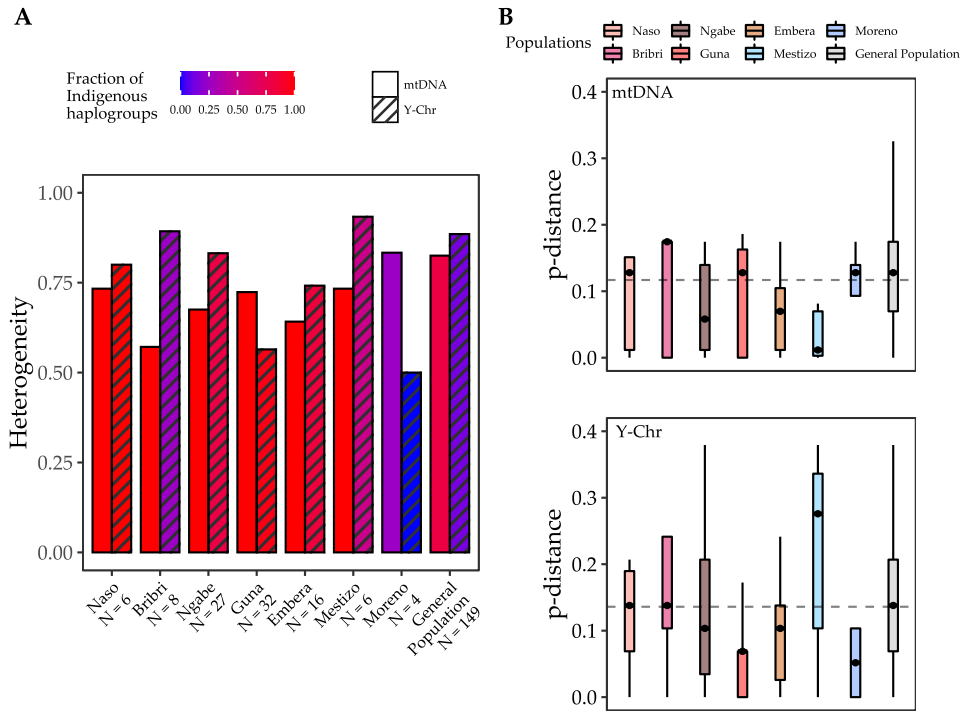


Figure 24. A) Heterogeneity based on haplogroup frequencies. Bars are shaded (from blue to red) according to the proportion of Indigenous lineages in each population. B) Intra-population uniparental genetic pairwise distances of different Panamanian groups computed for mtDNA and the Y-chromosome. Dashed lines represent the mean value of all distances for mtDNA (above, mean = 0.12) and for the Y chromosome (below, mean = 0.14). Adapted from Figure 4 in Rambaldi Migliore et al. [4] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Opposite trends are observed for the Guna group. The low level of Guna Y-chromosome heterogeneity is probably due to a strong bottleneck in the population size before contact and/or loss of male lineages in post-contact times. The Naso show similar values for the two. This observation is also confirmed by a much higher median value of genetic distance on the mtDNA gene pool for Bribri, Guna, and Moreno, thus testifying to a more diversified mtDNA gene pool with respect to the Y-chromosome one (although there could be a sample bias due to the sample size of Bribri and Moreno). Higher heterogeneity of the mtDNA pool is in agreement with the hypothesis of a very ancient IAm mtDNA legacy [5]. The high median value of the Y-chromosome heterogeneity estimates among the Mestizo may be due to the various allochthonous paternal contributions (both WEu and SAf) in post-contact times, although this pattern does not appear in the general population. To summarize the contribution of both uniparental lines to the gene pool of each individual, we computed the mean of Y-chromosome and mtDNA pairwise distances and used the resulting dissimilarity matrix to compute a MDS (Figure 25).

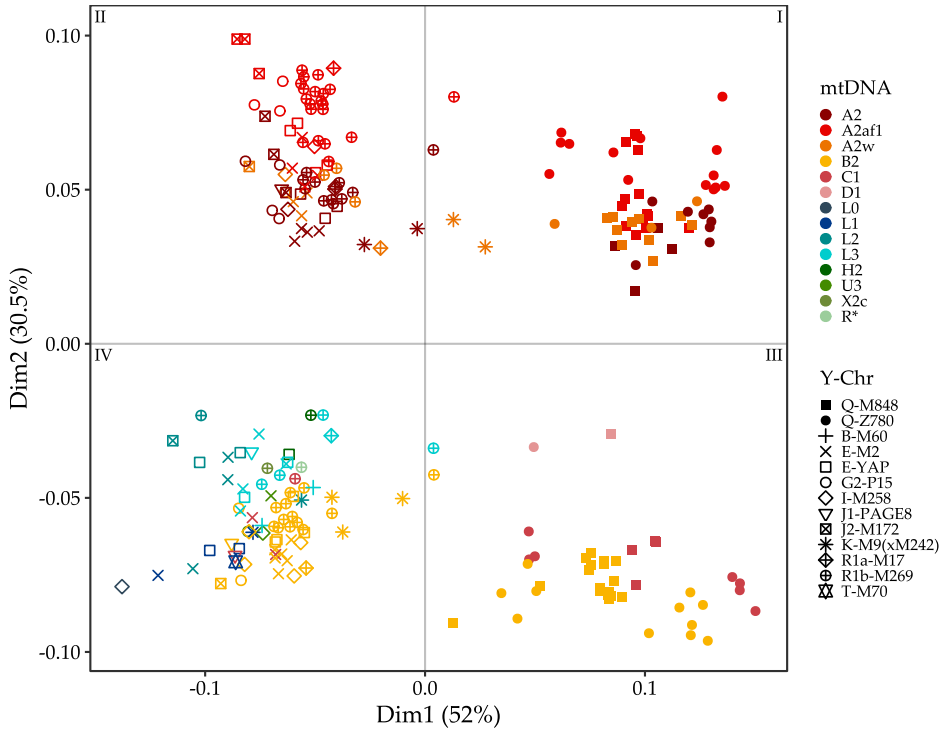


Figure 25. MDS plot computed on the mean of mtDNA and Y-chromosome pairwise distances. Adapted from Figure 6 in Rambaldi Migliore et al. [4] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The first dimension, accounting for 52% of the total variation, separates individuals with both paternal and maternal IAm haplogroups; this separation is mainly driven by the Y chromosomes, while Indigenous mtDNA Hgs form two main clusters along dimension two. The first clusters (quadrants I and II) include only individuals belonging to A2, whereas the clusters in the third and fourth quadrants enclose all individuals belonging to the other IAm lineages (B2, C1, and D1). WEu and SAf lineages characterize only mtDNAs falling in quadrant IV, instead. These findings indicate that the Indigenous uniparental lineages largely contributed to the current differentiation in Panama and further highlights the strong footprint of the mitochondrial Hgs A2af1 and A2w. A2af1 has been mostly found in the Isthmo-Colombian area, including pre-Hispanic individuals [5, 352, 359], and in a few individuals from Central Mexico [3], consistently with a probable origin from the north. The Y-chromosome lineage Q-M848* shows the same pattern of A2af1, with high frequencies across all Panamanian IAm populations, especially in the Guna (59.4% and 52.0%, respectively). This hints at the existence of an Isthmo-specific male source, yet to be identified, which could mirror the A2af1 lineage on the maternal side. Age estimates for A2af1 (15.82 ± 4.09 kya, [5]) and Q-M848 node (14.78 ± 0.02 kya, [363]) are in line with this hypothesis.

In conclusion, it is evident that both pre- and post-contact events contributed to shaping the uniparental gene pool of modern Panamanians. Before contact, the Indigenous groups probably remained isolated, forming three main clusters. The western Isthmian cluster is now represented by Bribri, Naso, and Ngäbe, despite some dissimilarities in the latter (i.e., the mitochondrial A2w and the Y-chromosome Q-M925*). This cluster is only clear on the maternal side. In the eastern Isthmus, the Guna preserve a specific Isthmian genome-wide component (see below and [5]) and show a major legacy of very ancient uniparental footprints, i.e. the previously defined mtDNA haplogroup A2af1 and the Y-chromosome paragroup Q-M848*. Conversely, the Emberá show traces of inputs from the south (e.g., the mtDNA haplogroup D1). These different footprints are still evident mostly in the mitochondrial gene pool because the maternal lines of Indigenous populations were only marginally involved in post-contact admixtures. We have confirmed a sex-biased introgression of only paternal non-Indigenous lineages into local communities, but our findings also updated previous assessments concerning sub-Saharan African genetic inputs. In fact, our new estimates on the general population, defined more accurately as formed by individuals not identifying themselves as members of any specific population group, contribute to solving an apparent discrepancy between genetic and historical data. This and other issues have been further investigated and directly tested through a genomic and diachronic comparison between pre- and post-contact individuals, as described hereafter.

8. Genome-wide analyses on modern and ancient individuals

Considering the new frontiers of archaeogenomics and the recently built aDNA facility in our Department, in parallel to the studies on mtDNA detailed above, I had the possibility to be involved in projects dealing with ancient and modern genome-wide data. It is well known that genomic analyses based only on mitogenomes have advantages (e.g., a very good molecular clock) and disadvantages (e.g., incomplete lineage sorting, drift, etc.). Certainly, this uniparental system can reconstruct only one of the several ancestries which can be instead identified by deciphering the “mosaic” of the entire genome. Therefore, a single genome can help reconstructing the complex history of a population and, even in population genetics, the analyses can be restricted to less individuals. In this view, a subset of the Panamanian individuals, previously analyzed for the uniparental systems, have been genotyped with high-density SNP arrays, while low-coverage ancient genomes have been obtained from pre-Hispanic and colonial individuals excavated in Panama City (chapter 8.1). Another, still ongoing, project moved further south along the double continent and focused on genome-wide data from an Indigenous Peruvian population (chapter 8.2).

8.1 Archaeogenomic distinctiveness of the Isthmo-Colombian area

An overview of Panama history has already been given in chapter 7.2.2. Considering its crucial geographic location, an archaeogenomic study of the Isthmus could reveal further hints on its past as well as on movements between North and South America. However, this region represented a gap in the genomic datasets of modern and ancient DNA sequences. This was mainly due to difficulties in collecting biological samples from Indigenous communities and to the bad preservation of ancient DNA in this tropical region. Nevertheless, in this work [5], we were able to produce 84 genome-wide profiles, genotyped with the Axiom™ Genome-Wide Human Origins 1 Array (see chapter 4.2), from five Indigenous (Bribri, Emberá Guna, Naso, Ngäbe) and two admixed (Moreno, Mestizo) groups currently living in Panama. Moreover, we shotgun-sequenced the first 12 low-coverage ancient genomes from the Isthmus. These ancient DNAs were extracted from an original collection of 20 human remains (13 pre-colonial and seven colonial) retrieved from seven different archeological excavations along the Pacific coast of Panama City. These data were compared with modern and ancient available genome-wide data from the Americas and selected worldwide populations that left a greater genomic impact on Indigenous Americans

during colonial times [66, 347, 364, 365].

To validate the aDNA data produced in this project, I investigated the possible modern human contamination in our sequence data using the three different methods detailed in chapter 6.2.3. These analyses revealed that two male individuals had mitochondrial contamination higher than 3%, while X chromosome contamination could not be estimated due to their low coverage. These two sequences were not used in downstream analyses, which provided novel and multidisciplinary information on the history of the Isthmian area.

We provided final answers to long-lasting anthropological and archaeological questions concerning the possible kinship relationships among individuals buried together in the sites of *Panamá Viejo* and *Coco del Mar*, in both of which one adult female skeleton was surrounded by male crania. The absence of biological relationships together with the long-time span of the different remains (~1000 years) allowed us to hypothesize that they were female seers with heads of prestigious enemies (obtained in warfare) that would have facilitated their access to their knowledge, as well as their ability to communicate with other worlds.

The major outcomes of this work were on the genetic history. ADMIXTURE analyses (Figure 26A) revealed distinctive genomic profiles which differentiate the gene pool of the Panamanian groups. The two admixed groups Moreno and Mestizo, as expected, revealed large proportions of their genomes not derived from Indigenous peoples of the Americas. This allochthonous contribution was also observed in individuals who self-identified as Indigenous and genealogically unadmixed but showing variable amounts of African and European ancestries in their genomes, with the lowest average values in the Guna and Ngäbe.

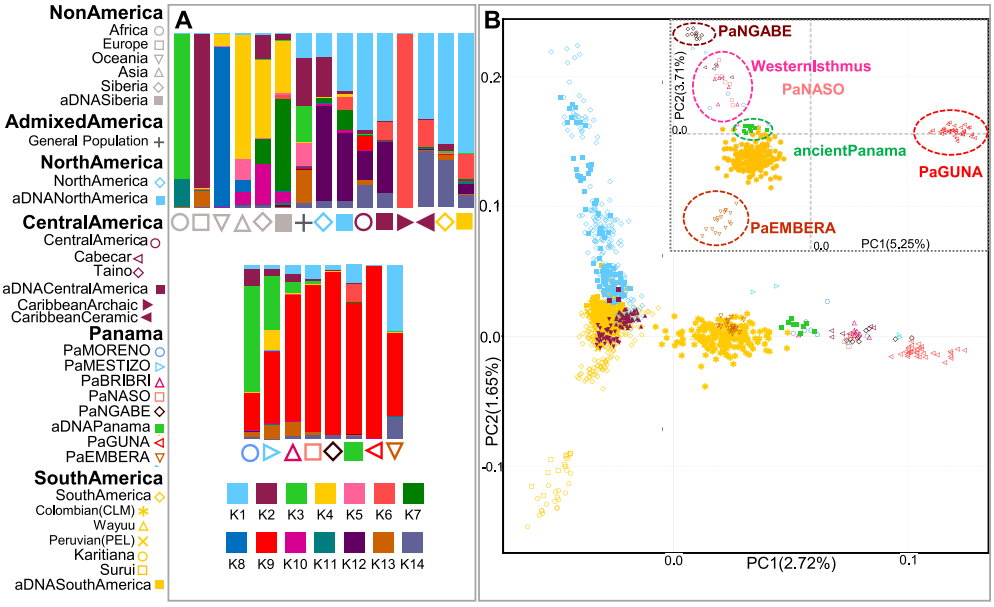


Figure 26. Overview of the genetic structure of ancient and modern Isthmian groups. A) ADMIXTURE plot for K = 14 (the K with the lowest cross-validation error from K1 to K20). Each bar shows the average ancestry proportion of individuals within the same group considering a modern worldwide dataset (1,560 individuals and 545,942 SNPs before pruning) plus American and Siberian ancient individuals (N = 341). (B) PCA analysis computed on individuals with more than 95% of Indigenous American ancestry (217 individuals and 534,569 SNPs). Ancient genomes (N = 341) and individuals with masked non-Indigenous ancestry (N = 417) were projected onto the modern variability. The inset shows a specific Isthmo-Colombian PCA. Adapted from Figure 2 in Capodiferro et al. [5] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

The modern and ancient Isthmian individuals are also characterized by a specific Indigenous component (Figure 26A) that probably drives the Isthmo-Colombian axis observed along the PC1 in the PCA analysis of Indigenous American groups (Figure 26B). We also identified a remarkable genomic structure within Panama, largely overlapping with past and present Indigenous groups (inset of Figure 26B). This implies that each group has its own genetic history. However, some groups (Bribri, Cabécar, Naso Djërdi and Ngäbe) also show relatedness, especially in the western Isthmian area, extended to southeastern Costa Rica. Fewer genetic similarities were identified between the Indigenous groups in eastern Panama (the Guna and Emberá), and between the Emberá and the pre-Hispanic Panamanians. These macro-groups seem to reflect the pre-colonial cultural areas, especially in the western Greater Chiriquí region, while the link in the eastern region, previously known as the Greater Darién, is less evident.

The present research had an impact that expands far beyond the Isthmian area. The ADMIXTURE analysis revealed the prevalence of one component (represented by the red color in Figure 26A) that is not found in other

Indigenous groups and makes the Panamanian genomes unique. A pre-Hispanic origin of the Isthmo-Colombian distinctiveness was suggested by the estimates of effective population size we obtained from the analysis of identical-by-descent (IBD) fragments. A reduction in the population size of the Panamanian groups probably started in pre-colonial times (~1 kya), thus before the average time of other Indigenous American populations. Moreover, shared IBD fragments among the Panamanian groups and other Indigenous American populations revealed ancient interactions within the Isthmo-Colombian area. However, based on the limited number and resolution of our genomic data, it was not possible to provide a more precise time frame and an accurate demographic reconstruction. Therefore, we employed the well-calibrated mtDNA molecular clock and my main contribution to this work was the analysis of modern and ancient mtDNAs from 80 (out of 84) present-day Panamanians (belonging to Indigenous lineages) and nine ancient pre-colonial individuals. The Bayesian phylogenetic tree (Figure 27) shows that the most represented haplogroups among ancient and modern Panamanian mitogenomes belong to the four main founding lineages (A2, B2, C1, and D1). Moreover, the comparison with published mitogenomes belonging to the same sub-haplogroups highlighted the presence of four Isthmian-specific sub-lineages (in black in Figure 27), the most represented one being A2af1, which has been dated to ~13 kya and confirms the ancient legacy of Isthmo-Colombian populations. Finally, the BSP based on Panamanian mitogenomes (inset of Figure 27) shows an increase in the effective population size starting ~10 kya in the Early Holocene.

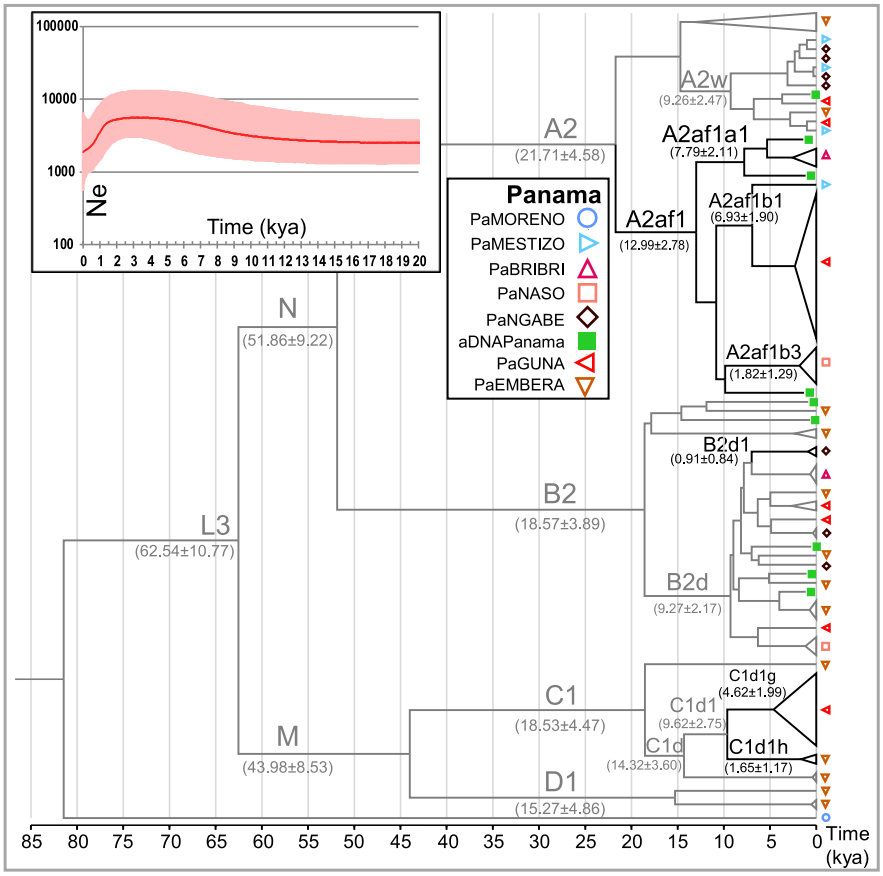


Figure 27. Bayesian phylogenetic tree of nine ancient and 80 modern mitogenomes from Panama belonging to Indigenous American founding Hgs. The tree was rooted on an L2c2 mitogenome from a ‘‘Moreno’’ individual. The Bayesian age (mean value with standard deviation) is shown for relevant branches. Black lines highlight Isthmo-Colombian specific branches. The inset shows the BSP, displaying changes in the effective N_e through time considering a generation time of 25 years. Adapted from Figure 5 in Capodiferro et al. [5] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

In conclusion, this work confirms that the ancient and modern Indigenous groups on the Isthmus of Panama share a common genetic history with other Indigenous peoples in North, Central and South America, but along a specific variation axis, clearly detectable within the America’s genetic landscape. On a continental scale, the comparison with available ancient and modern genomic data revealed a distinctive Isthmo-Colombian Indigenous genetic component. The ancestral origin of this genomic distinctiveness is summarized in the graphical abstract of our paper (Figure 28), where we tried to represent the different ancestries (with different colors) brought by the first Indigenous American groups (here also indicated as Native Americans, NA) into the Americas along different migration paths (dotted lines). The stars indicate, with different colors, admixture events that have been identified so

far by studying both modern and ancient individuals.

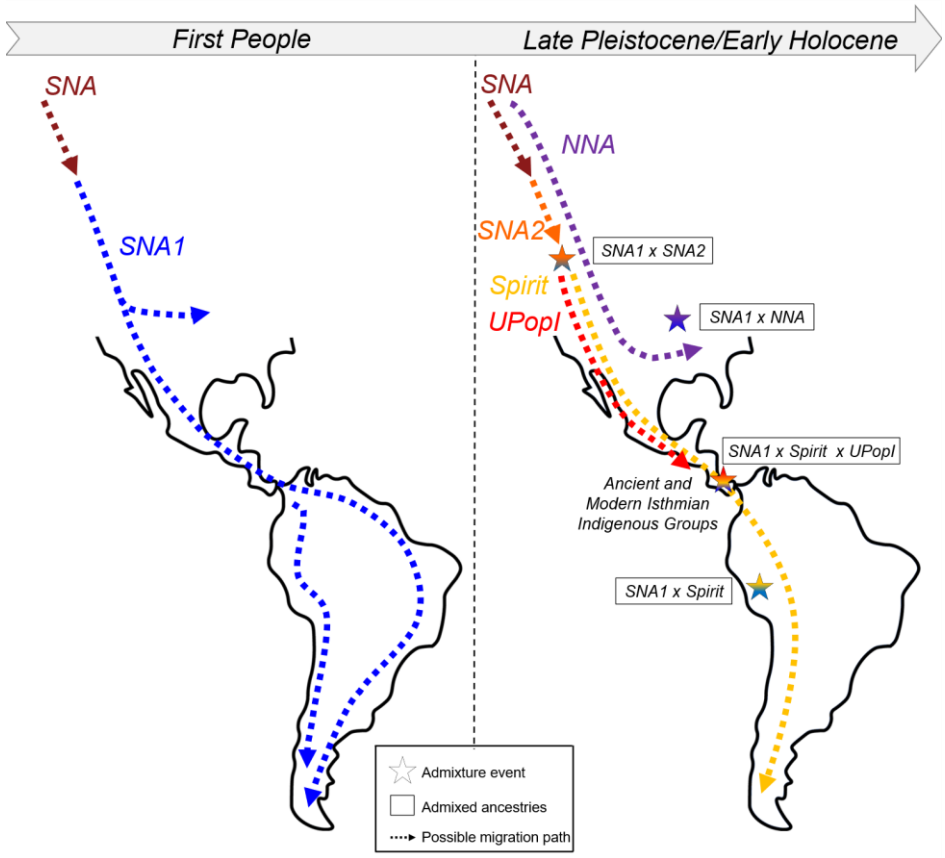


Figure 28. Ancestries and admixture events that shaped the archaeogenomic distinctiveness of the Isthmo-Colombian area. Adapted from the graphical abstract in Capodiferro et al. [5] under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Coming from Beringia, a land bridge connecting northern north America with eastern Siberia during the LGM (~25-18 kya), the first settlers left signs all over the double continent, spreading the so-called SNA1 (Southern Native American 1, blue lines in the left panel) ancestry, reaching the southern cone ~15 kya, as confirmed by archaeological evidence. However, recent studies of ancient and modern genomes describe a complex scenario prior to European contact with multiple migrations from Beringia, as initially suggested by mitochondrial DNA, as well as demographic spreads and admixture events along the two continents. The great majority of these additional ancestries (right panel) differentiated sometime between the late Pleistocene and the early Holocene (~15-8 kya), based on the earliest archaeological human evidence. One ancestry remained limited to northern north America (NNA, purple line) without crossing the Isthmus. Another one,

strongly related to a Pleistocene ancient individual from the Spirit Cave, Nevada (dated to ~10.9 kya), passed the land bridge and reached southern south America leaving strongest traces on the Pacific coast (yellow line). However, to fully explain the genetic variation of ancient and modern Panamanians, we need to consider an additional ancestry of northern American origin that parallel the Spirit Cave one, but which remained restricted to the Isthmian area (red line). We called this previously undescribed ancestry Unsampled Population of the Isthmus (UPopI), which is still unrepresented in ancient datasets probably because it was bound to now-submerged archeological sites on the Pacific coast of the Isthmus. Nevertheless, the genomes of the pre-Hispanic individuals from Panama City (i.e., from the archeological areas of *Panama Viejo* and *Coco del Mar*) attest to these events.

The genomic data we used in this study provide a starting point for future interdisciplinary studies on the Isthmo-Colombian crossroads. We are currently sequencing high-coverage genomes from present-day Panamanians and ancient (admixed) human remains dated to early colonial times. This will allow us to refine the region's genetic history with more statistical power and higher molecular resolution, by for example detailing the genetic patterns we found in the present study and to investigate the demographic changes of Indigenous population before, during, and after European contact.

8.2 A genome-wide survey of Ashaninka from Peru

The peculiar geophysical characteristics of South America created over time a perfect setting for the development of complex population dynamics, migrations, and adaptations, which targeted the attention of many scholars. I was involved in the so-called South American genome project, which aims to provide a genomic portrait of South America to infer and date a number of events concerning the history and pre-history of the human populations that, at different times, inhabited the subcontinent.

Among south American countries, Peru has been subject to several studies on modern Indigenous populations and ancient individuals [42, 208, 215, 218, 366, 367], which highlighted a genomic structure, often reflecting geographical features [218, 366]. Peruvian Amazonian communities show a higher genetic homogeneity in comparison to Andean and coastal groups, probably due to longer isolation periods [367]. The Ashaninka is the most numerous Indigenous groups of Peru, mostly living in the Pasco region between the eastern slope of the Andes and Yurua in the Ucayali region of western Amazonia. Ashaninkas speak a language that belongs to the Arawak family. Previous analyses on the uniparental systems showed the mtDNA

haplogroup D and Y-chromosome haplogroup Q as the most represented lineages, with low genetic input of non-Indigenous maternal lineages [368, 369]. To extend the molecular screening, we genotyped 51 Ashaninka individuals from the Pasco region with the Axiom™ Genome-Wide Human Origins 1 Array. After quality check, 44 Ashaninka genome-wide profiles were compared to a worldwide dataset of modern and ancient individuals, mainly from the Americas.

The Ashaninka shows an outlier position in the PCA at the continental level, with Caribbean, Amazonian, and other Peruvian populations as the closest genetic neighbors (data not shown). When considering the South America continent (Figure 29A), they are more closely related to groups from north Peru than to eastern Amazonians, particularly far from those living in the east and who experienced isolation periods (i.e., Surui and Karitiana). Allele frequency analyses (PCA and ADMIXTURE, Figure 29) also show that the 44 Ashaninka individuals cluster into two main homogeneous genetic groups. The most numerous, *Ashaninka1*, comprises a high percentage of individuals with less than 5% of non-Indigenous ancestry, while this percentage is higher in the second cluster, *Ashaninka2*. A third cluster, *Ashaninka3*, was also identified, but with a genetic profile different from the others. All three clusters have their own cline of variation in the PCA and show the highest percentages of a specific Indigenous component (in red) in the ADMIXTURE analysis (at K16, Figure 29B).

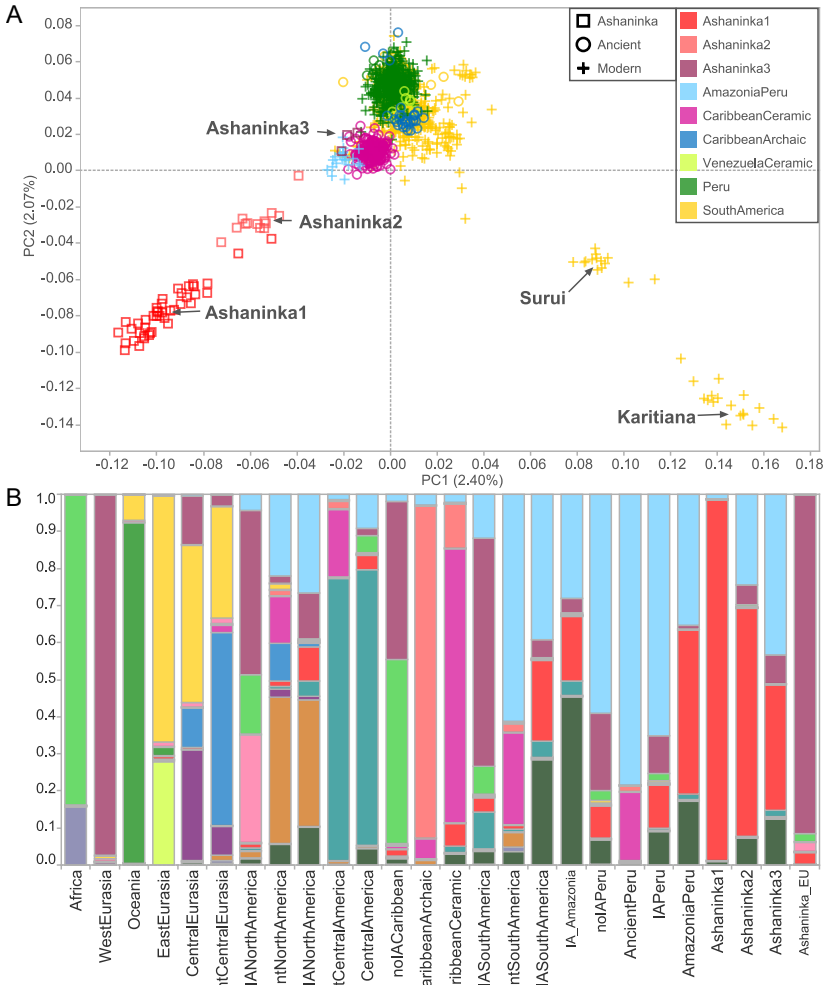


Figure 29. A) PCA computed only on South American modern and ancient Indigenous people, selected from the datasets of all individuals with more than 95% of Indigenous American ancestry (245 individuals and 409,084 SNPs), of IAm ancient genomes and individuals with masked non-Indigenous ancestry (N = 455) which were projected onto the modern variability. B) ADMIXTURE analyses of the modern dataset (1,604 individuals and 409,084 SNPs before pruning) and ancient individuals (N = 557) at K16 (the one with the lowest cross-validation error from K2 to K20). Unpublished Figure.

These findings confirm that various gene flows differentially impacted the Ashaninka group(s), as already suggested [218]. As evident from the PCA depicted in Figure 29A and from preliminary results from f_4-statistics, our analyses also support a previously hypothesized Arawak connection to the Caribbean [220]. Further analyses will be required to build a statistically significant model that could explain and summarize these findings. Multidisciplinary inputs, particularly from archaeology, anthropology, history, and linguistics, will be essential to interpret this genetic pattern and/or to formulate alternative hypotheses to be tested.

9. Conclusions and future perspectives

This thesis highlights some key features in the field of population genetics. First, mtDNA studies are still fundamental to build accurate phylogenies and for evolutionary reconstructions, with implications also in forensic genetics. However, this uniparental system can trace back only one ancestral (matrilineal, successful, and unbroken) path out of the thousands that contributed to a present-day genome. This complexity can be grasped by accompanying autosomal data to uniparental systems. Furthermore, the study of ancient DNA (known as palaeogenomics or archaeogenomics) has become an opportunity for a real-time investigation on the genetic past of populations. Finally, depending on the specific topic and/or objective, it is also crucial to accurately set the molecular screening at both a macro- and microgeographic scale.

The projects presented here combined one or more of these aspects and investigated the genetic histories of different human populations from Eurasia and the Americas. Three works were conducted exclusively on the mitochondrial DNA. In the first study, the diachronic comparison of 198 complete mtDNAs (selected from 545 modern mtDNA control regions) with 19 ancient mitogenomes from the Umbria region in Italy highlight a long and complex history of migrations from different sources and in different times, with genetic continuity between ancient and modern individuals in the eastern part of the region. The second research analyzed 2,420 modern mtDNAs (147 at the level of complete mitogenomes) from all over Mongolia. The results suggest: (i) continuous connections with East Asian neighboring populations until recent times, (ii) a link with post-glacial repopulation events from Western Eurasian refuges, and (iii) a connection with the path and the time frame of the Silk Route. In another macrogeographic work, a country-wide dataset of 2,021 mtDNAs from the present-day general population of Mexico revealed that the genetic impact of European conquest was modest in terms of maternal lineage introgression, with preservation of Indigenous haplogroups. Moreover, we show the importance of country-wide and regional databases at least for forensic genetic investigations.

The other research extended the analysis to the entire genome. In a still ongoing project on Peru, we are analyzing genome-wide autosomal data from 44 Ashaninka Indigenous individuals. This population shows an outlier behavior in the South American genomic landscape and at least two main genetic clusters can be identified. Moreover, peculiar connections with modern and ancient individuals from Central/South America emerged, especially with ancient Caribbean groups (dated to the Ceramic Age), a link that will need further investigation.

My main projects focused on the Isthmus of Panama and encompassed all

the (previously mentioned) main characteristics of an archaeogenomic population study. The characterization of the uniparental genetic history of 431 individuals from Indigenous and admixed groups and the general population shows that the local mtDNA gene pool (especially among the Indigenous populations) was marginally involved in post-contact admixtures, whereas the Indigenous Y chromosomes were differentially replaced mostly by West Eurasian lineages. Moreover, we provide new estimates of the sub-Saharan African contribution to a more accurately defined general population of Panama. In the second project, we increased the molecular resolution power. A total of 84 genome-wide profiles from Indigenous and admixed groups currently living in Panama were analyzed, together with the first reliable 12 low-coverage ancient genomes from this region. The results identify genomic structure within Panama, although some groups show relatedness, especially in the western regions, mirroring pre-contact cultural areas. We also describe for the first time a distinctive Isthmo-Colombian component in the Americas' genomic landscape, which derives from the admixture of different ancestries, including a still unsampled population that likely reached the Isthmus in the late Pleistocene, expanded during the Holocene, and left genomic traces up to present times.

My future work will build on the results of these two works on the Isthmus of Panama, trying to investigate two aspects: (i) the sex-biased gene flows from Europe and Africa into Indigenous American populations after contact, and (ii) the decline in population size of Indigenous groups before and during European expansion, and the subsequent demographic recovery. To this aim, the next steps of the Panama project will include whole genomes from present-day individuals and high-coverage ancient genomes from early colonial times. These data will be used to investigate gene flows and demographic changes with more statistical power and higher molecular resolution.

10. References

1. Modi A, Lancioni H, Cardinali I, Capodiferro MR, Rambaldi Migliore N, Hussein A, et al. The mitogenome portrait of Umbria in Central Italy as depicted by contemporary inhabitants and pre-Roman remains. Scientific Reports. 2020;10:10700.

2. Cardinali I, Bodner M, Capodiferro MR, Amory C, Rambaldi Migliore N, Gomez J E, et al. Mitochondrial DNA Footprints from Western Eurasia in Modern Mongolia. Frontiers in Genetics. 2022;:2749.

3. Bodner M, Perego UA, Gomez JE, Cerda-Flores RM, Rambaldi Migliore N, Woodward SR, et al. The mitochondrial DNA landscape of modern Mexico. Genes. 2021;12:1453.

4. Rambaldi Migliore N, Colombo G, Capodiferro MR, Mazzocchi L, Chero Osorio AM, Raveane A, et al. Weaving Mitochondrial DNA and Y-Chromosome Variation in the Panamanian Genetic Canvas. Genes. 2021;12:1921.

5. Capodiferro MR, Aram B, Raveane A, Rambaldi Migliore N, Colombo G, Ongaro L, et al. Archaeogenomic distinctiveness of the Isthmo-Colombian area. Cell. 2021;184:1706-1723.e24.

6. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics. 2010;19:R131–6.

7. Tautz D. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. DNA Fingerprinting: State of the Science. 1993;:21–8.

8. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nature Genetics. 2003;33:266–75.

9. Fisher RA. XXI.—On the dominance ratio. Proceedings of the royal society of Edinburgh. 1923;42:321–41.

10. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97.

11. Kimura M. The neutral theory of molecular evolution. 1983.

12. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;:624–6.

13. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1982;13:235–48.

14. Hartl DL, Clark AG. Principles of population genetics. 1997;116.

15. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proceedings of the National Academy of Sciences. 2016;113:5652–7.

16. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics. 2012;13:745–53.

17. Ho SY, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular Ecology. 2014;23:5947–65.

18. Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. Heredity. 2017;118:2–9.

19. Hirszfeld L, Hirszfeldowa H. Essai d'application des méthodes sérologiques au problème des races. 1919.

20. Harris H. C. Genetics of Man Enzyme polymorphisms in man. Proceedings of the Royal Society of London. Series B. Biological Sciences. 1966;164:298–310.

21. Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics. 1966;54:577.

22. Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science. 1949;110:543–8.

23. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. Journal of Experimental Medicine. 1944;79:137–58.

24. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953;171:740–1.

25. Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature. 1953;171:737–8.

26. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics. 1980;32:314.

27. Mullis KB, Faloona FA. [21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology. 1987;155:335–50.

28. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977;74:5463–7.

29. Yang A, Zhang W, Wang J, Yang K, Han Y, Zhang L. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Frontiers in Bioengineering and Biotechnology. 2020;8:1032.

30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;:860–921.

31. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291:1304–51.

32. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011;12:443–51.

33. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011;12:363–76.

34. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics. 2007;39:1251–5.

35. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.

36. Metzker ML. Sequencing technologies—the next generation. Nature Reviews Genetics. 2010;11:31–46.

37. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68.

38. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.

39. Bergström, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367.

40. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.

41. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nature Genetics. 2015;47:435–44.

42. Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, et al. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proceedings of the National Academy of Sciences. 2018;115:E6526–35.

43. Leitsalu L, Alavere H, Tammesoo M-L, Leego E, Metspalu A. Linking a population biobank with national health registries—the Estonian experience. Journal of Personalized Medicine. 2015;5:96–106.

44. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519:309–14.

45. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science. 2014;344:1280–5.

46. Pankratov V, Montinaro F, Kushniarevich A, Hudjashov G, Jay F, Saag L, et al. Differences in local population history at the finest level: the case of the Estonian population. European Journal of Human Genetics. 2020;28:1580–91.

47. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends in Genetics. 2018;34:666–81.

48. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. BioRxiv. 2021. https://doi.org/10.1101/2021.05.26.445798.

49. Gershman A, Sauria ME, Hook PW, Hoyt SJ, Razaghi R, Koren S, et al. Epigenetic patterns in a complete human genome. bioRxiv. 2021.

50. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. bioRxiv. 2021.

51. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757–62.

52. Gregory TR. Understanding Evolutionary Trees. Evolution: Education and Outreach. 2008;1:121–37.

53. Sharma A, Jaloree S, Thakur RS. Review of Clustering Methods: Toward Phylogenetic Tree Constructions. Proceedings of International Conference on Recent Advancement on Computer and Communication. 2018;:475–80.

54. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.

55. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. American Journal of Human Genetics. 2009;84:740–59.

56. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genetics. 2006;2:e190.

57. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

58. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19:1655–64.

59. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93.

60. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461:489–94.

61. Peter BM. Admixture, population structure, and F-statistics. Genetics. 2016;202:1485–501.

62. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods. 2012;9:179–81.

63. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genetics. 2012;8:e1002453.

64. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics. 2013;93:278–88.

65. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343:747–51.

66. Chacón-Duque J-C, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nature Communications. 2018;9:1–13.

67. Schiffels S, Wang K. MSMC and MSMC2: the multiple sequentially markovian coalescent. Methods in Molecular Biology. 2020;:147–66.

68. Kamm J, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association. 2020;115:1472–87.

69. Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation. 2009;30:E386–94.

70. Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. Fine-scale population structure and the era of next-generation sequencing. Human

Molecular Genetics. 2010;19:R221–6.

71. Kivisild T. Maternal ancestry and population history from whole mitochondrial genomes. Investigative Genetics. 2015;6:1–10.

72. Wilkins JF. Unraveling male and female histories from human genetic data. Current Opinion in Genetics & Development. 2006;16:611–7.

73. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nature Reviews Genetics. 2003;4:598–612.

74. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. Harvesting the fruit of the human mtDNA tree. Trends in Genetics. 2006;22:339–45.

75. Avise JC, Nelson WS, Bowen BW, Walker D. Phylogeography of colonially nesting seabirds, with special reference to global matrilineal patterns in the sooty tern (Sterna fuscata). Molecular Ecology. 2000;9:1783–92.

76. Jobling MA, Hurles M, Tyler-Smith C. Human evolutionary genetics: origins, peoples and disease. 2014.

77. van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome. Human Mutation. 2014;35:187–91.

78. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016;48:593–9.

79. Kurland C, Collins L, Penny D. Genomics and the irreducible nature of eukaryote cells. Science. 2006;312:1011–4.

80. Stewart JB, Larsson N-G. Keeping mtDNA in shape between generations. PLoS Genetics. 2014;10:e1004670.

81. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290:457–65.

82. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics. 1999;23:147–147.

83. Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, et al. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. International journal of Legal Medicine. 2004;118:294–306.

84. Hagström E, Freyer C, Battersby BJ, Stewart JB, Larsson N-G. No recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal germline. Nucleic Acids Research. 2013;42:1111–6.

85. Hutchison CA, Newbold JE, Potter SS, Edgell MH. Maternal inheritance of mammalian mitochondrial DNA. Nature. 1974;251:536–8.

86. Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, et al. The structure of human mitochondrial DNA variation. Journal of Molecular Evolution. 1991;33:543–55.

87. Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences. 1979;76:1967–71.

88. Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, et al. Asian affinities and continental radiation of the four founding Native American mtDNAs. American Journal of Human Genetics. 1993;53:563.

89. Torroni A, Miller JA, Moore LG, Zamudio S, Zhuang J, Droma T, et al. Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. American Journal of Physical Anthropology. 1994;93:189–99.

90. Torroni A, Lott MT, Cabell MF, Chen Y-S, Lavergne L, Wallace DC. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. American Journal of Human Genetics. 1994;55:760.

91. Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. American Journal of Human Genetics. 1995;57:133.

92. Dür A, Huber N, Parson W. Fine-Tuning Phylogenetic Alignment and Haplogrouping of mtDNA Sequences. International Journal of Molecular Sciences. 2021;22:5747.

93. Rebolledo-Jaramillo B, Su MS-W, Stoler N, McElhoe JA, Dickins B, Blankenberg D, et al. Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proceedings of the National Academy of Sciences. 2014;111:15474–9.

94. Frank A, Lobry J. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene. 1999;238:65–77.

95. Bandelt H-J, Kong Q-P, Richards M, Macaulay V. Estimation of mutation rates and coalescence times: some caveats. Human Mitochondrial DNA and the Evolution of Homo sapiens. 2006;:47–90.

96. Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. Current Biology. 2016;26:827–33.

97. Cerezo M, Achilli A, Olivieri A, Perego UA, Gomez-Carballa A, Brisighelli F, et al. Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Research. 2012;22:821–6.

98. Underhill P, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annual Review of Genetics. 2007;41.

99. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences. 2013;110:2223–7.

100. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 2005;308:1034–6.

101. Petraglia M, Korisettar R, Boivin N, Clarkson C, Ditchfield P, Jones S, et al. Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. Science. 2007;317:114–6.

102. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514:445–9.

103. Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H-J, et al. The archaeogenetics of Europe. Current Biology. 2010;20:R174–83.

104. Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Current Biology. 2010;20:R188–93.

105. O'Rourke DH, Raff JA. The human genetic history of the Americas: the final frontier. Current Biology. 2010;20:R202–7.

106. Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Kashani BH, et al. Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proceedings of the National Academy of Sciences. 2013;110:14308–13.

107. Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, Woodward SR, et al. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE. 2008;3:e1764.

108. Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E, et al. The Paleo-Indian entry into South America according to mitogenomes. Molecular Biology and Evolution. 2018;35:299–311.

109. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, et al. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Current Biology. 2009;19:1–8.

110. Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH, et al. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Research. 2010;20:1174–9.

111. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, et al. Beringian standstill and spread of Native American founders. PLoS ONE. 2007;2:e829.

112. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012;488:370–4.

113. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505:87–91.

114. Kayser M. The human genetic history of Oceania: near and remote views of dispersal. Current Biology. 2010;20:R194–201.

115. Tobler R, Rohrlach A, Soubrier J, Bover P, Llamas B, Tuke J, et al. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. Nature. 2017;544:180–4.

116. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics.

2001;29:217–22.

117. Liu Y, Mao X, Krause J, Fu Q. Insights into human history from the first decade of ancient human genomics. Science. 2021;373:1479–84.

118. Achilli A, Olivieri A, Semino O, Torroni A. Ancient human genomes—keys to understanding our past. Science. 2018;360:964–5.

119. Bergström A, Frantz L, Schmidt R, Ersmark E, Lebrasseur O, Girdland-Flink L, et al. Origins and genetic legacy of prehistoric dogs. Science. 2020;370:557–64.

120. Daly KG, Mattiangeli V, Hare AJ, Davoudi H, Fathi H, Doost SB, et al. Herded and hunted goat genomes from the dawn of domestication in the Zagros Mountains. Proceedings of the National Academy of Sciences. 2021;118.

121. Frantz LA, Bradley DG, Larson G, Orlando L. Animal domestication in the era of ancient genomics. Nature Reviews Genetics. 2020;21:449–60.

122. Kocher A, Papac L, Barquera R, Key FM, Spyrou MA, Hübler R, et al. Ten millennia of hepatitis B virus evolution. Science. 2021;374:182–8.

123. Librado P, Khan N, Fages A, Kusliy MA, Suchan T, Tonasso-Calvière L, et al. The origins and spread of domestic horses from the Western Eurasian steppes. Nature. 2021;598:634–40.

124. Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. Nature Reviews Genetics. 2019;20:323–40.

125. van der Valk T, Pečnerová P, Díez-del-Molino D, Bergström A, Oppenheimer J, Hartmann S, et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature. 2021;591:265–9.

126. Pääbo S. Molecular cloning of ancient Egyptian mummy DNA. Nature. 1985;314:644–5.

127. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences. 2007;104:14616–21.

128. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, et al. Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research. 2007;35:5717–28.

129. Hofreiter M, Jaenicke V, Serre D, Haeseler A von, Pääbo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Research. 2001;29:4793–9.

130. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362:709–15.

131. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.

132. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL, et al. Ligation bias in illumina next-generation DNA libraries: implications for sequencing

ancient genomes. PLoS ONE. 2013;8:e78575.

133. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research. 2010;38:e87–e87.

134. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370:20130624.

135. Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, et al. A time transect of exomes from a Native American population before and after European contact. Nature Communications. 2016;7:13175.

136. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS ONE. 2012;7:e34131.

137. Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, et al. High-Throughput DNA sequencing of ancient wood. Molecular Ecology. 2018;27:1138–54.

138. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. Royal Society Open Science. 2016;3:160239.

139. Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, et al. Genetic analyses from ancient DNA. Annual Review of Genetics. 2004;38:645–79.

140. Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biology. 2015;16:1–18.

141. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11.

142. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, De La Fuente C, Chan J, Spence JP, et al. Early human dispersals within the Americas. Science. 2018;362.

143. Vernot B, Zavala EI, Gómez-Olivencia A, Jacobs Z, Slon V, Mafessoni F, et al. Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. Science. 2021;372.

144. Zavala EI, Jacobs Z, Vernot B, Shunkov MV, Kozlikin MB, Derevianko AP, et al. Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave. Nature. 2021;1–5.

145. Smith RW, Non AL. Assessing the achievements and uncertain future of paleoepigenomics. Epigenomics. 2021.

146. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325:31–6.

147. Horai S. Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA. The Southeast Asian Journal of Tropical Medicine and Public Health. 1995;26:146–54.

148. Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000;408:708–13.

149. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991;253:1503–7.

150. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Nature. 2016;538:207–14.

151. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538:238–42.

152. Weidenreich F. Some problems dealing with ancient man. American Anthropologist. 1940;42:375–83.

153. Coon CS. The Rock Art of Africa. 1963.

154. Templeton AR. Genetics and recent human evolution. Evolution: International Journal of Organic Evolution. 2007;61:1507–19.

155. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053–60.

156. Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, et al. A high-coverage Neandertal genome from Chagyrskaya Cave. Proceedings of the National Academy of Sciences. 2020;117:15132–6.

157. Prüfer K, De Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358:655–8.

158. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.

159. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, et al. An early modern human from Romania with a recent Neanderthal ancestor. Nature. 2015;524:216–9.

160. Hajdinjak M, Mafessoni F, Skov L, Vernot B, Hübner A, Fu Q, et al. Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. Nature. 2021;592:253–7.

161. Kim BY, Lohmueller KE. Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. American Journal of Human Genetics. 2015;96:454–61.

162. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;343:1017–21.

163. Vernot B, Akey JM. Complex history of admixture between modern humans and Neandertals. American Journal of Human Genetics. 2015;96:448–53.

164. Skoglund P, Reich D. A genomic view of the peopling of the Americas. Current Opinion in Genetics & Development. 2016;41:27–35.

165. Larena M, McKenna J, Sanchez-Quinto F, Bernhardsson C, Ebeo C, Reyes R, et al. Philippine Ayta possess the highest level of Denisovan ancestry in the world. Current Biology. 2021;31:4219-4230. e10.

166. Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature. 2018;561:113–6.

167. Chan EK, Timmermann A, Baldi BF, Moore AE, Lyons RJ, Lee S-S, et al. Human origins in a southern African palaeo-wetland and first migrations. Nature. 2019;575:185–9.

168. Lipson M, Ribot I, Mallick S, Rohland N, Olalde I, Adamski N, et al. Ancient West African foragers in the context of African population history. Nature. 2020;577.

169. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing prehistoric African population structure. Cell. 2017;171:59-71. e21.

170. Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature. 2017;546:289–92.

171. Vidal CM, Lane CS, Asrat A, Barfod DN, Mark DF, Tomlinson EL, et al. Age of the oldest known Homo sapiens from eastern Africa. Nature. 2022;601:579–83.

172. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358:652–5.

173. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Current Biology. 2005;15:R159–60.

174. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences. 2005;102:15942–7.

175. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541:302–10.

176. DeGiorgio M, Jakobsson M, Rosenberg NA. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proceedings of the National Academy of Sciences. 2009;106:16057–62.

177. Llamas B, Rada XR, Collen E. Ancient DNA helps trace the peopling of the world. The Biochemist. 2020;42:18–22.

178. Prendergast ME, Lipson M, Sawchuk EA, Olalde I, Ogola CA, Rohland N, et al. Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. Science. 2019;365.

179. Wang K, Goldstein S, Bleasdale M, Clist B, Bostoen K, Bakwa-Lufu P, et al. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. Science Advances. 2020;6:eaaz0183.

180. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A,

et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334:94–8.

181. Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, et al. The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell. 2021.

182. Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, et al. 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. Current Biology. 2017;27:3202-3208. e9.

183. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, et al. The population history of northeastern Siberia since the Pleistocene. Nature. 2019;570:182–8.

184. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534:200–5.

185. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, et al. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346:1113–8.

186. Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of bronze age Eurasia. Nature. 2015;522:167–72.

187. O'Connell JF, Allen J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. Journal of Archaeological Science. 2015;56:73–84.

188. Bergström A, Nagle N, Chen Y, McCarthy S, Pollard MO, Ayub Q, et al. Deep roots for Aboriginal Australian Y chromosomes. Current Biology. 2016;26:809–13.

189. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. Nature. 2016;538:510–3.

190. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, et al. Population turnover in remote Oceania shortly after initial settlement. Current Biology. 2018;28:1157-1165. e7.

191. Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, et al. Language continuity despite population replacement in Remote Oceania. Nature Ecology & Evolution. 2018;2:731–40.

192. Pugach I, Hübner A, Hung H, Meyer M, Carson MT, Stoneking M. Ancient DNA from Guam and the peopling of the Pacific. Proceedings of the National Academy of Sciences. 2021;118.

193. Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, et al. Native American gene flow into Polynesia predating Easter Island settlement. Nature. 2020;583:572–7.

194. Yang MA, Fu Q. Insights into modern human prehistory using ancient genomes. Trends in Genetics. 2018;34:184–96.

195. Prüfer K, Posth C, Yu H, Stoessel A, Spyrou MA, Deviese T, et al. A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. Nature Ecology & Evolution. 2021;5:820–5.

196. Hublin J-J, Sirakov N, Aldeias V, Bailey S, Bard E, Delvigne V, et al. Initial Upper

Palaeolithic Homo sapiens from Bacho Kiro Cave, Bulgaria. Nature. 2020;581:299–302.

197. Skoglund P, Mathieson I. Ancient genomics of modern humans: the first decade. Annual Review of Genomics and Human Genetics. 2018;19:381–404.

198. Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, et al. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science. 2014;344:747–50.

199. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555:190–6.

200. Ardelean CF, Becerra-Valdivia L, Pedersen MW, Schwenninger J-L, Oviatt CG, Macías-Quintero JI, et al. Evidence of human occupation in Mexico around the Last Glacial Maximum. Nature. 2020;584:87–92.

201. Becerra-Valdivia L, Higham T. The timing and effect of the earliest human arrivals in North America. Nature. 2020;584:93–7.

202. Dillehay TD, Goodbred S, Pino M, Sánchez VFV, Tham TR, Adovasio J, et al. Simple technologies and diverse food strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru. Science Advances. 2017; 3(5):e1602778.

203. Battaglia V, Grugni V, Perego UA, Angerhofer N, Gomez-Palmieri JE, Woodward SR, et al. The first peopling of South America: new evidence from Y-chromosome haplogroup Q. PLoS ONE. 2013;8:e71390.

204. Dulik MC, Owings AC, Gaieski JB, Vilar MG, Andre A, Lennie C, et al. Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan-and Eskimoan-speaking populations. Proceedings of the National Academy of Sciences. 2012;109:8471–6.

205. Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, et al. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American Journal of Human Genetics. 2008;82:583–92.

206. Kashani BH, Perego UA, Olivieri A, Angerhofer N, Gandini F, Carossa V, et al. Mitochondrial haplogroup C4c: A rare lineage entering America through the ice-free corridor? American Journal of Physical Anthropology. 2012;147:35–9.

207. Willerslev E, Meltzer DJ. Peopling of the Americas as inferred from ancient genomics. Nature. 2021;594:356–64.

208. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, et al. Reconstructing the deep population history of Central and South America. Cell. 2018;175:1185-1197. e22.

209. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;349.

210. Yu H, Spyrou MA, Karapetian M, Shnaider S, Radzevičiūtė R, Nägele K, et al. Paleolithic to Bronze Age Siberians reveal connections with first Americans and across Eurasia. Cell. 2020;181:1232-1245. e20.

211. Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, et al. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature. 2018;553:203–7.

212. Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Science Advances. 2016;2:e1501385.

213. Pinotti T, Bergström A, Geppert M, Bawn M, Ohasi D, Shi W, et al. Y chromosome sequences reveal a short Beringian Standstill, rapid expansion, and early population structure of Native American founders. Current Biology. 2019;29:149-157. e3.

214. Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, et al. Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. Proceedings of the National Academy of Sciences. 2017;114:4093–8.

215. Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, et al. The genetic prehistory of the Andean highlands 7000 years BP though European contact. Science Advances. 2018;4:eaau4921.

216. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506:225–9.

217. Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, et al. Ancient human parallel lineages within North America contributed to a coastal expansion. Science. 2018;360:1024–7.

218. Nakatsuka N, Lazaridis I, Barbieri C, Skoglund P, Rohland N, Mallick S, et al. A paleogenomic reconstruction of the deep population history of the Andes. Cell. 2020;181:1131-1145. e21.

219. Nakatsuka N, Luisi P, Motti JM, Salemme M, Santiago F, del Campo MD, et al. Ancient genomes in South Patagonia reveal population movements associated with technological shifts and geography. Nature Communications. 2020;11:1–12.

220. Fernandes DM, Sirak KA, Ringbauer H, Sedig J, Rohland N, Cheronet O, et al. A genetic history of the pre-contact Caribbean. Nature. 2021;590:103–10.

221. Nägele K, Posth C, Orbegozo MI, De Armas YC, Godoy STH, Herrera UMG, et al. Genomic insights into the early peopling of the Caribbean. Science. 2020;369:456–60.

222. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015;525:104–8.

223. Fehren-Schmitz L, Harkins KM, Jarman CL. Paleogenomic investigations of human remains from Rapa Nui. American Journal of Physical Antrhopology. 2017;162:178–178.

224. Moreno-Mayar JV, Rasmussen S, Seguin-Orlando A, Rasmussen M, Liang M, Flåm ST, et al. Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans. Current Biology. 2014;24:2518–25.

225. Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, et al. Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. American Journal of Human Genetics. 1997;60:1107.

226. Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, Rengo C, et al. A Signal, from Human mtDNA, of Postglacial Recolonization in Europe. The American Journal of Human Genetics. 2001;69:844–52.

227. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Research. 2016;44:W58–63.

228. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. 2010.

229. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015;516:517.

230. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.

231. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

232. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4.

233. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20:1297–303.

234. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular Biology and Evolution. 2017;34:3299–302.

235. Nei M. Molecular evolutionary genetics. 1987.

236. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105:437–60.

237. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2021. 2021.

238. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution. 2018;35:1547.

239. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. Journal of Statistical Software. 2008;25:1–18.

240. Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. American Journal of Physical Anthropology. 2018;165:238–55.

241. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Computational Biology. 2019;15:e1006650.

242. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics. 2007;81:559–75.

243. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources. 2015;15:1179–91.

244. Rosenberg NA. distruct: a program for the graphical display of population structure. Molecular Ecology Notes. 2004;4:137–8.

245. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences. 2013;110:15758–63.

246. Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, et al. Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. PLoS One. 2015;10:e0129102.

247. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. Genome Biology. 2016;17:1–14.

248. Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE. 2010;5:e14004.

249. Modi A, Tassi F, Susca RR, Vai S, Rizzi E, De Bellis G, et al. Complete mitochondrial sequences from Mesolithic Sardinia. Scientific Reports. 2017;7:1–10.

250. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

251. Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Current Biology. 2013;23:553–9.

252. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:1–13.

253. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008;134:416–26.

254. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002;30:3059–66.

255. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013;30:772–80.

256. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;7:457–72.

257. Moreno-Mayar JV, Korneliussen TS, Dalal J, Renaud G, Albrechtsen A, Nielsen R, et al. A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. Bioinformatics. 2020;36:828–41.

258. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52.

259. Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, et al. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. Science. 2019;366:708–14.

260. Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, et al. Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS ONE. 2013;8:e65441.

261. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, et al. An overview of the genetic structure within the Italian population from genome-wide data. PLoS ONE. 2012; 7(9)::e43759.

262. Pereira JB, Costa MD, Vieira D, Pala M, Bamford L, Harich N, et al. Reconciling evidence from ancient and contemporary genomes: a major source for the European Neolithic within Mediterranean Europe. Proceedings of the Royal Society B: Biological Sciences. 2017;284:20161976.

263. Raveane A, Aneli S, Montinaro F, Athanasiadis G, Barlera S, Birolo G, et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. Science Advances. 2019;5:eaaw3492.

264. Sarno S, Boattini A, Carta M, Ferri G, Alù M, Yao DY, et al. An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of sicily and southern Italy. PLoS ONE. 2014;9:e96074.

265. Brisighelli F, Álvarez-Iglesias V, Fondevila M, Blanco-Verea A, Carracedo A, Pascali VL, et al. Uniparental markers of contemporary Italian population reveals details on its pre-Roman heritage. PLoS ONE. 2012;7:e50794.

266. Grugni V, Raveane A, Mattioli F, Battaglia V, Sala C, Toniolo D, et al. Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective. Annals of Human Biology. 2018;45:44–56.

267. Parolo S, Lisa A, Gentilini D, Di Blasio AM, Barlera S, Nicolis EB, et al. Characterization of the biological processes shaping the genetic structure of the Italian population. BMC Genetics. 2015;16:1–11.

268. Sazzini M, Ruscone GAG, Giuliani C, Sarno S, Quagliariello A, De Fanti S, et al. Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. Scientific Reports. 2016;6:1–11.

269. Vai S, Ghirotto S, Pilli E, Tassi F, Lari M, Rizzi E, et al. Genealogical relationships between early medieval and modern inhabitants of Piedmont. PLoS ONE. 2015;10:e0116801.

270. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, et al. Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. American Journal of Human Genetics. 2007;80:759–68.

271. Ghirotto S, Tassi F, Fumagalli E, Colonna V, Sandionigi A, Lari M, et al. Origins and evolution of the Etruscans' mtDNA. PLoS ONE. 2013;8:e55519.

272. Gómez-Carballa A, Pardo-Seco J, Amigo J, Martinón-Torres F, Salas A. Mitogenomes from The 1000 Genome Project reveal new Near Eastern features in present-day Tuscans. PLoS One. 2015;10:e0119242.

273. Guimaraes S, Ghirotto S, Benazzo A, Milani L, Lari M, Pilli E, et al. Genealogical discontinuities among Etruscan, Medieval, and contemporary Tuscans. Molecular Biology and Evolution. 2009;26:2157–66.

274. Leonardi M, Sandionigi A, Conzato A, Vai S, Lari M, Tassi F, et al. The female ancestor's tale: Long-term matrilineal continuity in a nonisolated region of Tuscany. American Journal of Physical Anthropology. 2018;167:497–506.

275. Pardo-Seco J, Gómez-Carballa A, Amigo J, Martinón-Torres F, Salas A. A genome-wide study of modern-day Tuscans: revisiting Herodotus's theory on the origin of the Etruscans. PLoS ONE. 2014;9:e105920.

276. Serventi P, Panicucci C, Bodega R, De Fanti S, Sarno S, Fondevila Alvarez M, et al. Iron Age Italic population genetics: the Piceni from Novilara (8th–7th century BC). Annals of Human Biology. 2018;45:34–43.

277. Tassi F, Ghirotto S, Caramelli D, Barbujani G. Genetic evidence does not support an Etruscan origin in Anatolia. American Journal of Physical Anthropology. 2013;152:11–8.

278. Galiberti. Il Paleolitico inferiore in Italia. 1982.

279. Bradley G. Ancient Umbria: state, culture, and identity in central Italy from the Iron Age to the Augustan era. 2000.

280. Bonomi-Ponzi L. La necropoli plestina di Colfiorito di Foligno. 1997.

281. Ebenesersdóttir SS, Sandoval-Velasco M, Gunnarsdóttir ED, Jagadeesan A, Guðmundsdóttir VB, Thordardóttir EL, et al. Ancient genomes from Iceland reveal the making of a human population. Science. 2018;360:1028–32.

282. Juras A, Makarowicz P, Chyleński M, Ehler E, Malmström H, Krzewińska M, et al. Mitochondrial genomes from Bronze Age Poland reveal genetic continuity from the Late Neolithic and additional genetic affinities with the steppe populations. American Journal of Physical Anthropology. 2020;172:176–88.

283. Knipper C, Mittnik A, Massy K, Kociumaka C, Kucukkalipci I, Maus M, et al. Female exogamy and gene pool diversification at the transition from the Final Neolithic to the Early Bronze Age in central Europe. Proceedings of the National Academy of Sciences. 2017;114:10083–8.

284. Krause-Kyora B, Nutsua M, Boehme L, Pierini F, Pedersen DD, Kornell S-C, et al. Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. Nature Communications. 2018;9:1–11.

285. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018;555:197–203.

286. Neparáczki E, Kocsy K, Tóth GE, Maróti Z, Kalmár T, Bihari P, et al. Revising mtDNA haplotypes of the ancient Hungarian conquerors with next generation

sequencing. PLoS ONE. 2017;12:e0174886.

287. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science. 2019;363:1230–4.

288. Sazzini M, Abondio P, Sarno S, Gnecchi-Ruscone GA, Ragno M, Giuliani C, et al. Genomic history of the Italian population recapitulates key evolutionary dynamics of both Continental and Southern Europeans. BMC Biology. 2020;18:51.

289. Kovalev AA, Erdenebaatar D. Discovery of new cultures of the Bronze Age in Mongolia according to the data obtained by the International Central Asian Archaeological Expedition. Current Archaeological Research in Mongolia. 2009;:149–70.

290. Wilkin S, Ventresca Miller A, Fernandes R, Spengler R, Taylor WT-T, Brown DR, et al. Dairying enabled Early Bronze Age Yamnaya steppe expansions. Nature. 2021;598:629–33.

291. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nature Genetics. 2018;50:1696–704.

292. Cavalli-Sforza L, Menozzi P, Piazza A. The history and geography of human genes. Princeton University Press, Princeton. 1994.

293. Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Martínez-Arias R, et al. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. American Journal of Human Genetics. 1998;63:1824–38.

294. Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, et al. The genetic history of admixture across inner Eurasia. Nature Ecology & Evolution. 2019;3:966–76.

295. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, et al. The formation of human populations in South and Central Asia. Science. 2019;365.

296. Ning C, Zheng H-X, Zhang F, Wu S, Li C, Zhao Y, et al. Ancient Mitochondrial Genomes Reveal Extensive Genetic Influence of the Steppe Pastoralists in Western Xinjiang. Frontiers in Genetics. 2021;:740167.

297. Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, et al. The complex admixture history and recent southern origins of Siberian populations. Molecular Biology and Evolution. 2016;33:1777–95.

298. Yang L, Tan S, Yu H, Zheng B, Qiao E, Dong Y, et al. Gene admixture in ethnic populations in upper part of Silk Road revealed by mtDNA polymorphism. Science in China Series C: Life Sciences. 2008;51:435–44.

299. Yao Y-G, Kong Q-P, Wang C-Y, Zhu C-L, Zhang Y-P. Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in china. Molecular Biology and Evolution. 2004;21:2265–80.

300. Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genetics. 2015;11:e1005068.

301. Jeong C, Wang K, Wilkin S, Taylor WTT, Miller BK, Bemmann JH, et al. A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. Cell. 2020;183:890-904. e29.

302. Wang C-C, Yeh H-Y, Popov AN, Zhang H-Q, Matsumura H, Sirak K, et al. Genomic insights into the formation of human populations in East Asia. Nature. 2021;591:413–9.

303. Köstenbauer J. Surgical wisdom and Genghis Khan's Pax Mongolica. ANZ Journal of Surgery. 2017;87:116–20.

304. Rogers LL, Honeychurch W, Amartuvshin C, Kaestle FA. U5a1 mitochondrial DNA haplotype identified in Eneolithic skeleton from Shatar Chuluu, Mongolia. Human Biology. 2020;91:213–23.

305. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, 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. American Journal of Human Genetics. 2004;75:910–8.

306. An C-B, Chen F-H, Barton L. Holocene environmental changes in Mongolia: a review. Global and Planetary Change. 2008;63:283–9.

307. Orkhonselenge A, Komatsu G, Uuganzaya M. Middle to late Holocene sedimentation dynamics and paleoclimatic conditions in the Lake Ulaan basin, southern Mongolia. Géomorphologie: relief, processus, environnement. 2018;24:351–63.

308. Chatters JC, Kennett DJ, Asmerom Y, Kemp BM, Polyak V, Blank AN, et al. Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. Science. 2014;344:750–4.

309. González-Martín A, Gorostiza A, Regalado-Liu L, Arroyo-Peña S, Tirado S, Nuño-Arana I, et al. Demographic history of indigenous populations in Mesoamerica based on mtDNA sequence data. PLoS ONE. 2015;10:e0131791.

310. Morales-Arce AY, Hofman CA, Duggan AT, Benfer AK, Katzenberg MA, McCafferty G, et al. Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico. Scientific Reports. 2017;7:1–13.

311. Sandoval K, Buentello-Malo L, Penaloza-Espinosa R, Avelino H, Salas A, Calafell F, et al. Linguistic and maternal genetic diversity are not correlated in Native Mexicans. Human Genetics. 2009;126:521–31.

312. González-Sobrino BZ, Pintado-Cortina AP, Sebastián-Medina L, Morales-Mandujano F, Contreras AV, Aguilar YE, et al. Genetic diversity and differentiation in urban and indigenous populations of Mexico: Patterns of mitochondrial DNA and Y-chromosome lineages. Biodemography and Social Biology. 2016;62:53–72.

313. Gorostiza A, Acunha-Alonzo V, Regalado-Liu L, Tirado S, Granados J, Sámano D, et al. Reconstructing the history of Mesoamerican populations through the study of the mitochondrial DNA control region. 2012.

314. Guardado-Estrada M, Juarez-Torres E, Medina-Martinez I, Wegier A, Macías A, Gomez G, et al. A great diversity of Amerindian mitochondrial DNA ancestry is present in the Mexican mestizo population. Journal of Human Genetics. 2009;54:695–705.

315. Kemp BM, González-Oliver A, Malhi RS, Monroe C, Schroeder KB, McDonough J, et al. Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. Proceedings of the National Academy of Sciences. 2010;107:6759–64.

316. Gojobori J, Mizuno F, Wang L, Onishi K, Granados J, Gomez-Trejo C, et al. mtDNA diversity of the Zapotec in Mexico suggests a population decline long before the first contact with Europeans. Journal of Human Genetics. 2015;60:557–9.

317. González-Oliver A, Pineda-Vázquez D, Garfias-Morales E, De La Cruz-Laina I, Medrano-González L, Márquez-Morfín L, et al. Genetic overview of the Maya populations: Mitochondrial DNA haplogroups. Human Biology. 2019;90:281–300.

318. Green LD, Derr JN, Knight A. mtDNA affinities of the peoples of North-Central Mexico. American Journal of Human Genetics. 2000;66:989–98.

319. Martínez-Cortés G, Salazar-Flores J, Haro-Guerrero J, Rubi-Castellanos R, Velarde-Félix JS, Muñoz-Valle JF, et al. Maternal admixture and population structure in Mexican–Mestizos based on mtDNA haplogroups. American Journal of Physical Anthropology. 2013;151:526–37.

320. Vidal O, Brusca R-C. Mexico's biocultural diversity in peril. Revista de Biología Tropical. 2020;68:669–91.

321. Kumar S, Bellis C, Zlojutro M, Melton PE, Blangero J, Curran JE. Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. BMC Evolutionary Biology. 2011;11:1–17.

322. Campos-Sánchez R, Barrantes R, Silva S, Escamilla M, Ontiveros A, Nicolini H, et al. Genetic structure analysis of three Hispanic populations from Costa Rica, Mexico, and the southwestern United States using Y-chromosome STR markers and mtDNA sequences. Human Biology. 2006;78:551–63.

323. Watkins WS, Xing J, Huff C, Witherspoon DJ, Zhang Y, Perego UA, et al. Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. BMC Genetics. 2012;13:1–14.

324. Mizuno F, Kumagai M, Kurosaki K, Hayashi M, Sugiyama S, Ueda S, et al. Imputation approach for deducing a complete mitogenome sequence from low-depth-coverage next-generation sequencing data: application to ancient remains from the Moon Pyramid, Mexico. Journal of Human Genetics. 2017;62:631–5.

325. Parson W, Dür A. EMPOP—a forensic mtDNA database. Forensic Science International: Genetics. 2007;1:88–92.

326. Bonilla C, Gutiérrez G, Parra EJ, Kline C, Shriver MD. Admixture analysis of a rural population of the state of Guerrero, Mexico. American Journal of Physical Anthropology. 2005;128:861–9.

327. Rosas RCB, Mercado Sesma A, Hernández Ortega L, Hernandez Gonzalez L, Vega Avalos J, Arreola Cruz AA. The utility of genomic public databases to mitochondrial haplotyping in contemporary Mestizo population of Mexican origin. Mitochondrial DNA Part A. 2019;30:567–72.

328. Rangel-Villalobos H, Muñoz-Valle J, González-Martín A, Gorostiza A, Magaña M, Páez-Riberos L. Genetic admixture, relatedness, and structure patterns among Mexican populations revealed by the Y-chromosome. American Journal of Physical

Anthropology. 2008;135:448–61.

329. Lopopolo M, Børsting C, Pereira V, Morling N. A study of the peopling of Greenland using next generation sequencing of complete mitochondrial genomes. American Journal of Physical Anthropology. 2016;161:698–704.

330. Schurr TG. The Peopling of the New World: Perspectives from Molecular Anthropology. Annual Review of Anthropology. 2004;33:551–83.

331. Brenner CH. Fundamental problem of forensic mathematics—the evidential value of a rare haplotype. Forensic Science International: Genetics. 2010;4:281–91.

332. Cooke R. Prehistory of native Americans on the Central American land bridge: Colonization, dispersal, and divergence. Journal of Archaeological Research. 2005;13:129–87.

333. Cooke RG, Sánchez HLA, Smith-Guzmán N, Lara-Kraudy A. Panamá prehispánico. Nueva historia general de Panamá. Panamá: Editora Novo Art, S.A; 2019.

334. Hernández Mora I, Martín JG, Aram B. The first Cathedral on America's Pacific coast. Historical Archaeology. 2021;55:219–37.

335. Cooke R, Ranere A, Pearson G, Dickau R. Radiocarbon chronology of early human settlement on the Isthmus of Panama (13,000–7000 BP) with comments on cultural affinities, environments, subsistence, and technological change. Quaternary International. 2013;301:3–22.

336. Ranere AJ, Cooke RG. Late glacial and Early Holocene migrations, and Middle Holocene settlement on the lower isthmian land-bridge. Quaternary International. 2021;578:20–34.

337. Piperno DR. The origins of plant cultivation and domestication in the New World tropics: patterns, process, and new developments. Current Anthropology. 2011;52:S453–70.

338. Piperno DR. Prehistoric human occupation and impacts on Neotropical forest landscapes during the Late Pleistocene and Early/Middle Holocene. Tropical Rainforest Responses to Climatic Change. 2007;:193–218.

339. Ulloa FC. La Gran Chiriquí: Una Historia cada vez más profunda. 2017.

340. O'Connor L, Muysken P. The native languages of South America: Origins, development, typology. 2014.

341. Romoli K. Los de la lengua de Cueva: Los grupos indígenas del istmo oriental en la época de la conquista española. 1987.

342. Cook ND. The Columbian Exchange. In: Bentley JH, Wiesner-Hanks ME, Subrahmanyam S, editors. The Cambridge World History: Volume 6: The Construction of a Global World, 1400–1800 CE. Cambridge: Cambridge University Press; 2015. p. 103–34.

343. Castillero-Calvo A. Conquista, evangelización y resistencia. 2017.

344. Castillero-Calvo A. Nueva Historia General de Panamá. Vol. I, tomo 1. Comisión del Bicentenario de la Fundación de Panamá Alcaldía de Panamá. 2019.

345. Fortes-Lima C, Verdu P. Anthropological genetics perspectives on the

transatlantic slave trade. Human Molecular Genetics. 2021;30:R79–87.

346. Ongaro L, Molinaro L, Flores R, Marnetto D, Capodiferro MR, Alarcón-Riquelme ME, et al. Evaluating the impact of sex-biased genetic admixture in the Americas through the analysis of haplotype data. Genes. 2021;12:1580.

347. Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, et al. The genomic impact of European colonization of the Americas. Current Biology. 2019;29:3974-3986. e4.

348. Carvajal-Carmona LG, Soto ID, Pineda N, Ortíz-Barrientos D, Duque C, Ospina-Duque J, et al. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. American Journal of Human Genetics. 2000;67:1287–95.

349. Melton PE, Baldi NF, Barrantes R, Crawford MH. Microevolution, migration, and the population structure of five Amerindian populations from Nicaragua and Costa Rica. American Journal of Human Biology. 2013;25:480–90.

350. Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortiz-Barrientos D, et al. Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre-and post-Columbian patterns of gene flow in South America. American Journal of Human Genetics. 2000;67:1277–86.

351. Seielstad M. Asymmetries in the maternal and paternal genetic histories of Colombian populations. American Journal of Human Genetics. 2000;67:1062.

352. Perego UA, Lancioni H, Tribaldos M, Angerhofer N, Ekins JE, Olivieri A, et al. Decrypting the mitochondrial gene pool of modern Panamanians. PloS ONE. 2012;7:e38337.

353. Grugni V, Battaglia V, Perego UA, Raveane A, Lancioni H, Olivieri A, et al. Exploring the Y chromosomal ancestry of modern Panamanians. PLoS ONE. 2015;10:e0144223.

354. Newson L, Minchin S. From Capture to Sale: The Portuguese Slave Trade to Spanish South America in the Early Seventeenth Century. Brill; 2007.

355. Vega Franco ML. El tráfico de esclavos con América: Asientos de Grillo y Lomelín (1663-1674): 297. Sevilla: Consejo Superior de Investigaciones Cientificas; 1984.

356. Behar DM, Van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A “Copernican” reassessment of the human mitochondrial DNA tree from its root. American Journal of Human Genetics. 2012;90:675–84.

357. Just RS, Irwin JA, Parson W. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Science International: Genetics. 2015;18:131–9.

358. Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, et al. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investigative Genetics. 2014;5:1–17.

359. Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, et al. Improved calibration of the human mitochondrial clock using ancient genomes. Molecular Biology and Evolution. 2014;31:2780–92.

360. Söchtig J, Álvarez-Iglesias V, Mosquera-Miguel A, Gelabert-Besada M, Gómez-Carballa A, Salas A. Genomic insights on the ethno-history of the Maya and the 'Ladinos' from Guatemala. BMC Genomics. 2015;16:1–18.

361. Alves-Silva J, da Silva Santos M, Guimarães PEM, Ferreira ACS, Bandelt H-J, Pena SDJ, et al. The Ancestry of Brazilian mtDNA Lineages. American Journal of Human Genetics. 2000;67:444–61.

362. Bisso-Machado R, Fagundes NJ. Uniparental genetic markers in Native Americans: A summary of all available data from ancient and contemporary populations. American Journal of Physical Anthropology. 2021;176:445–58.

363. Colombo G, Traverso L, Mazzocchi L, Grugni V, Rambaldi Migliore N, Capodiferro MR, et al. Overview of the Americas' First Peopling from a Patrilineal Perspective: New Evidence from the Southern Continent. Genes. 2022;13:220.

364. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic insights into the ancestry and demographic history of South America. PLoS Genetics. 2015;11:e1005602.

365. Montinaro F, Busby GB, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. Nature Communications. 2015;6:1–7.

366. Barbieri C, Barquera R, Arias L, Sandoval JR, Acosta O, Zurita C, et al. The current genomic landscape of western south America: Andes, amazonia, and Pacific coast. Molecular Biology and Evolution. 2019;36:2698–713.

367. Borda V, Alvim I, Mendes M, Silva-Carvalho C, Soares-Souza GB, Leal TP, et al. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proceedings of the National Academy of Sciences. 2020;117:32557–65.

368. Simão F, Xavier C, Tineo D, Carvalho E, Parson W, Gusmão L. The maternal inheritance of the Ashaninka native group from Peru. Forensic Science International: Genetics Supplement Series. 2019;7:135–7.

369. Tineo D, Loiola S, Paredes F, Noli L, Amaya Y, Simão F, et al. Genetic characterization of 27 Y-STR loci in the native population of Ashaninka from Peru. Forensic Science International: Genetics Supplement Series. 2015;5:e220–2.

11. List of publications

During my PhD, I coauthored seven papers, including six original research publications (#1-4 and #6-7) and one review article (#5), as listed hereafter:

1. G. Colombo*, L. Traverso*, L. Mazzocchi, V. Grugni, **N. Rambaldi Migliore**, M.R. Capodiferro, G. Lombardo, R. Flores, M. Karmin, S. Rootsi, L. Ferretti, A. Olivieri, A. Torroni, R. Martiniano, A. Achilli, A. Raveane*, O. Semino. Overview of the Americas' first peopling from a patrilineal perspective: new evidence from the Southern Continent. *Genes* 2022, 13(2): 220.
   DOI: https://doi.org/10.3390/genes13020220.

2. I. Cardinali*, M. Bodner*, M.R. Capodiferro*, C. Amory, **N. Rambaldi Migliore**, E.J. Gomez, E. Myagmar, T. Dashzeveg, F. Carano, S.R. Woodward, W. Parson, U.A. Perego, H. Lancioni, A. Achilli. Mitochondrial DNA footprints from Western Eurasia in modern Mongolia. *Frontiers in Genetics* 2022, 2:819337.
   DOI: https://doi.org/10.3389/fgene.2021.819337.

3. **N. Rambaldi Migliore***, G. Colombo*, M.R. Capodiferro, L. Mazzocchi, A.M. Chero Osorio, A. Raveane, M. Tribaldos, U.A. Perego, T. Mendizábal, A. García Montón, G. Lombardo, V. Grugni, M. Garofalo, L. Ferretti, C. Cereda, S. Gagliardi, R. Cooke, N. Smith-Guzmán, A. Olivieri, B. Aram, A. Torroni, J. Motta, O. Semino*, A. Achilli*. Weaving mitochondrial DNA and Y-chromosome variation in the Panamanian genetic canvas. *Genes* 2021, 12(12):1921.
   DOI: https://doi.org/10.3390/genes12121921.

4. M. Bodner*, U.A. Perego*, J.E. Gomez. R.M. Cerda Flores, **N. Rambaldi Migliore**, S.R. Woodward, W. Parson, A. Achilli. The mitochondrial DNA landscape of modern Mexico. *Genes* 2021, 12(9):1453.
   DOI: https://doi.org/10.3390/genes12091453.

5. S. Dato, P. Crocco, **N. Rambaldi Migliore**, F. Lescai. Omics in a digital world: the role of bioinformatics in providing new insights into human aging. *Frontiers in Genetics* 2021, 12:689824.
   DOI: https://doi.org/10.3389/fgene.2021.689824.

6. M.R. Capodiferro, B. Aram, A. Raveane, **N. Rambaldi Migliore**, G. Colombo, L. Ongaro, J. Rivera, T. Mendizábal, I. Hernández-Mora, M. Tribaldos, U.A. Perego, H. Li, C.L. Scheib, A. Modi, A. Gòmez-Carballa, V. Grugni, G. Lombardo, G. Hellenthal, J.M. Pascale, F. Bertolini, G. Grieco, C. Cereda, M. Lari, D. Caramelli, L. Pagani, M. Metspalu, R. Friedrich, C. Knipper, A. Olivieri, A. Salas, R. Cooke, F. Montinaro, J. Motta, A. Torroni, J.G. Martín, O. Semino, R.S. Malhi, A. Achilli. Archaeogenomic Distinctiveness of the Isthmo-Colombian Area. *Cell* 2021, 184(7):1706-1723.
DOI: https://doi.org/10.1016/j.cell.2021.02.040.

7. A. Modi*, H. Lancioni*, I. Cardinali*, M.R. Capodiferro*, **N. Rambaldi Migliore**, A. Hussein, C. Strobl, M. Bodner, L. Schnaller, C. Xavier, E. Rizzi, L. Bonomi Ponzi, S. Vai, A. Raveane, B. Cavadas, O. Semino, A. Torroni, A. Olivieri, M. Lari, L. Pereira, W. Parson, D. Caramelli, A. Achilli. The mitogenome portrait of Umbria in Central Italy as depicted by contemporary inhabitants and pre-Roman remains. *Scientific Reports* 2020, 10:10700.
DOI: https://doi.org/10.1038/s41598-020-67445-0.

* These authors contributed equally