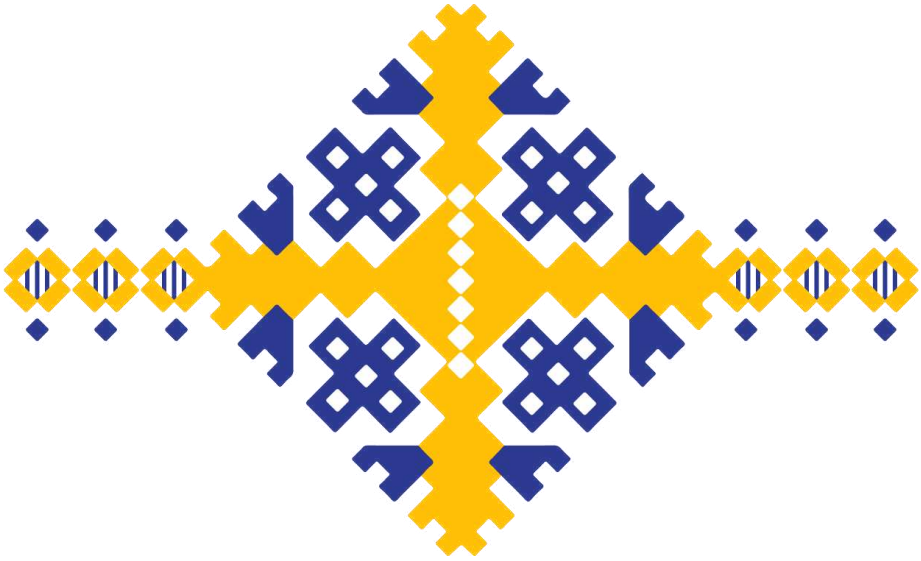
UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Archaeogenomic analyses of modern and ancient individuals from present-day Ukraine



Nataliia Kozak

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXVII – A.A. 2021-2024

Cover illustration created by Mariia Kulymova

UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

Archaeogenomic analyses of modern and ancient individuals from present-day Ukraine

Nataliia Kozak

Supervised by Prof. Alessandro Achilli

Dottorato di Ricerca in

Genetica, Biologia Molecolare e Cellulare

Ciclo XXXVII – A.A. 2021-2024

У ці нестерпні і буремні часи, не дивлячись на відстань,

усі мої думки з тобою, моя Батьківщино.

Ми разом повернемо нашу історію,

Ми вистоїмо!

Table of Contents

1. Abstract

Present-day Ukraine has been a crossroads of peoples and cultures since the Paleolithic, due to its strategic location on the northern shore of the Black Sea, between Eastern Europe and the Pontic-Caspian steppe. It served as one of the glacial refuges (the East European Plain) where populations survived during the Last Glacial Maximum and later repopulated Europe. Ukraine was also the cradle of the Slavic and Crimean Tatar civilizations which significantly contributed to the gene pool of Eastern Europe. Its strategic location, fertile lands, and favourable climate attracted migrations, settlements, and conflicts. These have resulted in a high number of admixture events that have layered the genetic history of Ukrainians, making it valuable to untangle this complex genetic variation for reconstructing the genetic landscape of Europe and Western Eurasia.

The ultimate goal of this work of thesis was to investigate the genetic variability of the Donetsk oblast population within the context of Western Eurasia for reconstructing its origins and genetic history. Through separate and combined analyses of genomic data from modern and ancient individuals living in the territory of present-day Ukraine, this dissertation aims to add another piece of information to the complex genetic landscape of Europe.

The first part of the thesis is focused on the current population of the Donetsk oblast, which is located in the region of the Great Eurasian Steppe, on the border between the nomadic pastoralists and more sedentary communities. This has contributed to the complex history of admixture in the area. The most recent migratory processes in this region took place during the 19th and 20th centuries in the context of urbanisation and social modernization. Nowadays, Donetsk oblast borders with the Russian Federation, and still a large part of the region is occupied by Russian forces since 2014; therefore, it is of particular interest to reveal the events of migration and gene flow with the surrounding regions, which left their own footprint in the formation of the local population. The gene pool of the modern population of Donetsk oblast (in year 1999) has been described through the study of 91 complete mtDNA sequences, complemented by genome-wide data from 45 individuals to provide both matrilineal and biparental perspectives.

Phylogenetic analysis of the complete mitogenomes revealed extensive mitochondrial DNA variation, with 80 lineages predominantly of western Eurasian origin (96.7%), plus two East Asian haplogroups (D4 and G3; 3.3%). Age estimates suggest that some haplogroups (i.e., D, U5, U8, I, and J2) have been present on the territory of Ukraine since pre-glacial times. Bayesian skyline analysis indicated two population growth phases: one during the Paleolithic (44–38 kya), coinciding with the arrival of modern humans in Europe, and another during the Neolithic (11–6 kya), linked to

agricultural expansion. High haplotype diversity (0.999) and low nucleotide diversity (0.0018) suggest a population bottleneck followed by rapid growth, probably associated with Bronze Age Yamna expansions. Genome-wide data from Donetsk oblast individuals clustered with Eastern Slavic groups (Russians, Belarusians) and Northern Europeans (England, Lithuania) in PCA, reflecting historical connections. Admixture analysis revealed two main components: one from Neolithic farmers and another from Eneolithic steppe pastoralists.

The second part of this work contributed to the archaeogenomic reconstruction of the history of Ukraine by generating mitochondrial and low-coverage genomic data via shotgun sequencing of ancient samples found on the territory of modern Ukraine. This evidence lays the genomic groundwork for a diachronic study of modern inhabitants and ancient individuals from Ukraine to provide insights into the ancestral origins of early settlers and later migrants.

We analysed seven ancient individuals belonging to three different cultures: two from Yamna, three from Catacomb, and two from Scythian. Radiocarbon (^{14}C) dating confirmed their cultural affiliation. Three ancient DNA sequences reached a sufficient quality for further analysis: a Catacomb male and two Scythian females. Kinship analyses revealed that these two ancient Scythian genomes represent the same individual. When compared to other ancient genomes and projected onto the modern genetic landscape of Western Eurasia, the Catacomb genome consistently clusters with other Catacomb individuals as well as with others available for the Yamna and Corded Ware cultures, which shared similar territories. This “steppe” group lies between Northern European and Turkic ethnic groups, supporting existing theories about the origins of the Kurgan cultures. The merged Scythian genome falls between previously reported Scythians from Kazakhstan and Ukraine on one axis, and between Northern/Eastern Europe and the Caucasus/Middle East on the other axis. This positioning suggests significant admixture due to the extensive nomadic lifestyle of the Scythians. The closest ancient genomes are from the Sarmatians, who succeeded the Scythians around the 3rd century BCE. This connection points to genetic continuity in the area, which is also suggested by the diachronic comparison of ancient and modern mitochondrial haplogroups. The mitochondrial haplogroup H1, found in the Catacomb individual and in nine modern Ukrainians, is frequently observed in Slavic populations nowadays. The Scythian individual belongs to haplogroup D5a2a2, which is a sister clade of the D4 branch detected in the Donetsk oblast.

In brief, this work of thesis has contributed to a better understanding of the genetic history of Europe by adding an important piece of information to the complex genetic landscape of Eastern Europe through the concurrent

analysis of mitogenomes and genomic (low-coverage) data from modern and ancient individuals of present-day Ukraine.

2. Abbreviations

1KGP 1000 Genomes Project

A adenine

aDNA ancient DNA

ANGSD Analysis of Next Generation Sequencing Data

bam binary alignment map

BCE Before Common Era

BEAST Bayesian Evolutionary Analysis by Sampling Trees

bp(s) base pair(s)

BSP Bayesian Skyline Plot

BWA Burrows-Wheeler Aligner

C cytosine

CE Common Era

CI confidence interval

CRS Cambridge Reference Sequence

cv cross-validation

ddNTPs dideoxynucleotides

D-Loop displacement loop

dNTPs deoxynucleotides

DNA Deoxyribonucleic acid

dsDNA double-stranded DNA

D-style Damgaard-style cementum sampling

FTDNA Family Tree DNA

G guanine

Gb Gigabases

Hg(s) haplogroup(s)

HGDP Human Genome Diversity Project

HGP Human Genome Project

HMM Hidden Markov Model

HO The Human Origins array

HS High Sensitivity

Kb(s) kilobase(s)

kDa kiloDalton

kya thousand years ago

LD Linkage Disequilibrium

LGM Last Glacial Maximum

LR PCR long-range PCR

MAFFT Multiple Alignment Fast Fourier Transform

Mb Megabase

MCMC Markov chain Monte Carlo

MDE minimally destructive extraction method

MDS Multidimensional Scaling

ML Maximum Likelihood

MP Maximum Parsimony

MRCA Most Recent Common Ancestor

MSY Male Specific Region of the Y chromosome

mtDNA mitochondrial DNA

Ne effective population size

NFW nuclease-free water

NGS Next Generation Sequencing

NJ Neighbor-joining

np(s) nucleotide position(s)

PC(s) Principal component(s)

PCA Principal Component Analysis

PCR Polymerase chain reaction

PNK Polynucleotide Kinase

rCRS revised Cambridge Reference Sequence

READ Relationship Estimation from Ancient DNA

rRNA ribosomal RNA

RSB Resuspension Buffer

SBS sequencing by synthesis

SNP(s) Single Nucleotide Polymorphism(s)

STR(s) Short Tandem Repeat(s)

T thymine

TMA terminal maternal ancestor

TPA terminal paternal ancestor

tRNA transfer RNA

U uracil

UDG uracil-DNA glycosylase

UGI uracil glycosylase inhibitor

USER uracil-specific excision reagent

UV ultraviolet

vcf variant calling format

WGS Whole Genome Sequencing

WTR Whole Tooth Root

ya years ago

3. Introduction

3.1 Population genetics overview

It is important to know the answers to questions about people's own origins and lay more evidence on the history written in textbooks. The contribution of population genetics in answering these questions has become fundamental in the last decades (*Torroni* 2006, *Gasparre* 2020). This branch of genetics investigates genetic variation in different populations and how they evolved during time and space. Its particular focus stays on the distributions and frequencies of alleles. Changes in the occurrence of particular genotypes can be a result of natural selection, mutations, migrations or genetic drift. All these factors, but particularly natural selection and genetic drift, are able to change the survival of organisms, therefore the passage of the allele through generations (*Cavalli-Sforza 2003*).

Natural selection can explain to us the evolution of specific genes or the preservation of some genotypes, such as the high frequency of heterozygotes for the gene of sickle cell anemia in some regions of Africa, where malaria is widespread, which gives opportunity to natives to survive despite of the plasmodium parasite presence. Another clear example is lactose intolerance in many human populations, such as East and South Asians, native Americans and indigenous Australians, some Arabs, and West Africans. In the past none of these groups had a significant contribution of dairy farming in the diets, therefore, there was no selective pressure to maintain lactase production into adulthood (*Ingram 2009*). Population genetics can also provide the answers about events shaping the genetic pool of populations, e.g. isolations, migrations, founder effects or bottlenecks.

What is the source of all those events and what gives living organisms flexibility to adapt and survive? The key answer here is the complex organisation of the hereditary material, Deoxyribonucleic acid (DNA), which serves as the only information storage and passage source for almost all living things. Despite all the mechanisms of preservation and repair of this "molecule of life", mutations occur in every form of life. Mutations are an essential driver for evolution, allowing to survive and adapt to some niches in competitive environments.

Speaking about human beings, DNA is organised in two sets of chromosomes, which are inherited from the parents and therefore, every gene for an individual can be represented only by two alleles. Each allele has its own particular location on the chromosome which is called *locus*. The pair of alleles from the same *locus* is called a genotype (*Krebs 2018*).

For all organisms characterised by sexual reproduction another source of variation is the recombination of parental DNAs during meiosis, through crossing-over and segregation. This random shuffling between homologous chromosomes creates new combinations of traits inherited from ancestors. When the frequency of certain alleles is higher or lower than expected by chance, it is called linkage disequilibrium (LD). There are some exceptions for particular DNA molecules, which do not have recombination and are transmitted from generation to generation, and evolve only through accumulation of mutations. In humans those molecules are the mitochondrial DNA (mtDNA) and the non-recombining region of the Y chromosome (*Watson 2014*).

Mutations occur from time to time in every organism. The measure which estimates the probability of mutations occurring in a genome each generation is called the mutation rate. For humans this rate is approximately 1.2×10^{-8} mutations per base pair (bp) per generation (*Scally 2012*). The most common type of mutation is the substitution of one nucleotide with another one and it is called single nucleotide polymorphism (SNP). To be classified as SNPs, the mutation rate must reach at least one percent frequency in a population. If it occurs less frequently, such alterations are called private mutations. Another type of genomic variation includes insertions and deletions (indels) and short tandem repeats (STRs), which usually involve larger segments of genetic material, ranging from one to six base pairs (*Mullaney 2010*). Lastly, larger structural changes can occur, such as deletions, duplications, insertions, inversions, and various types of translocations (*Pritchard 2023*).

There are several statistical models which are trying to predict the direction of evolution driven by genetic drift. The first fundamental model was proposed independently by two geneticists, R. Fisher and S. Wright, at the beginning of the 20th century. The "Wright-Fisher model" assumes that the population has a fixed number of individuals (N) overtime, with non-overlapping generations, random mating and no selection towards or against any alleles. Frequencies of the alleles change over generations due to random sampling process, leading to genetic drift. This causes the alleles frequency to fluctuate randomly, sometimes resulting in allele fixation or loss (*Fisher 1923, Wright 1931*). Later this model was upgraded to make it fitting to the real populations (*Ewens 2004, Kimura 1968*).

Another important model developed to understand the evolutionary process is the coalescent theory, which is derived from the Wright-Fisher model and was implemented by (*Kingman 1982*). This is a backward-in-time theory that traces an ancestry of an individual allele back to the most recent common ancestor (MRCA). The main idea here is that a pair of alleles merge to one ancestral allele in the past. The estimated coalescent time can be used for estimation of selection, genetic drift and population size changes. The coalescent model is based on Markov process (stochastic process, where the

probability of transitioning to the next state depends only on the current state, not on the sequence of events that preceded it) and it says that the rate of merging between two lineages is inversely proportional to the population size. If population size equals N_e (effective population size), the expected coalescence of two alleles will occur in $2N_e$ generations. This process is done to all sampled alleles of different individuals until a single common ancestor is found, forming a genealogical tree.

Another crucial concept of molecular biology, the molecular clock, states that "DNA and protein sequences evolve at a rate that is relatively constant over time and among different organisms" (*Ho 2008*). It means that evolution of different species is proportional to time when these species diverged and goes with the same frequency. This theory based on empirical observations of E. Zuckerkandi and L. Pauling is a useful instrument for estimation of timescales of evolution. Studying fossils, they observed that in hemoglobin proteins the number of differences in amino acids of different species correlated with their divergence times (*Ho 2008*). Back in the 1960s, the first theoretical backing by M. Kimura suggested that mutation rate is fixed and that mutations do not affect organisms' evolutionary fitness. This theory, of course, simplifies the complexity of biological organisms and population mechanics, and nowadays it is considered as a strict molecular clock. To "relax" the clock and make it closer to reality, two more relaxed-clock models were developed, one of which assumes that the rate fluctuates over time and between different species, with variation mostly around an average number. The second one assumes that the pace of molecular evolution is linked to other biological traits that also experience change, and permits the evolutionary rate to "evolve" over time. Calibration of molecular clocks must be done with independent data, i.e. fossil records or known biological or geological events, then the clock can be used for estimation of divergence times for other organisms or events.

All these fundamental theories are still used in population genetics by curious scientists who are trying to figure out the timeline of life. These concepts and their new extinctions give us a chance to trace events as migrations and to find divergence back to MRCA.

3.2 Population genetics technologies

Nowadays we have additional opportunities given by the most up-to-date methods of genetics and archaeogenetics to study with higher efficiency migrations and the demographic history of humanity.

Following the discovery of DNA structure, it came to the comprehension that, to advance in phenotype understanding, knowledge about the sequence of the nucleic acid molecule is crucial. A breakthrough happened with the development of polymerase chain reaction (PCR) technology, which allowed to obtain high copy numbers of targeted DNA sequences (*Mullis 1987*). Sequencing of DNA began in the 1970s with Sanger sequencing (*Sanger 1977*), which was dominant for a few decades. In this technology, the ddNTPs, or chemically modified deoxynucleotides, which randomly attach to the synthesised DNA chain, are substituted for regular dNTPs in this sequencing approach, because they cannot extend the elongation due to the lack of 3' hydroxyl group. This method is called “chain termination method”, since it allows obtaining different length fragments with dye-labelled ddNTPs in the ends. In this way, by making a comparison between the different chains, by capillary electrophoresis and fluorescence detection it is possible to trace the order of all bases and so, make the sequencing (*Metzker 2010, Heather 2016*).

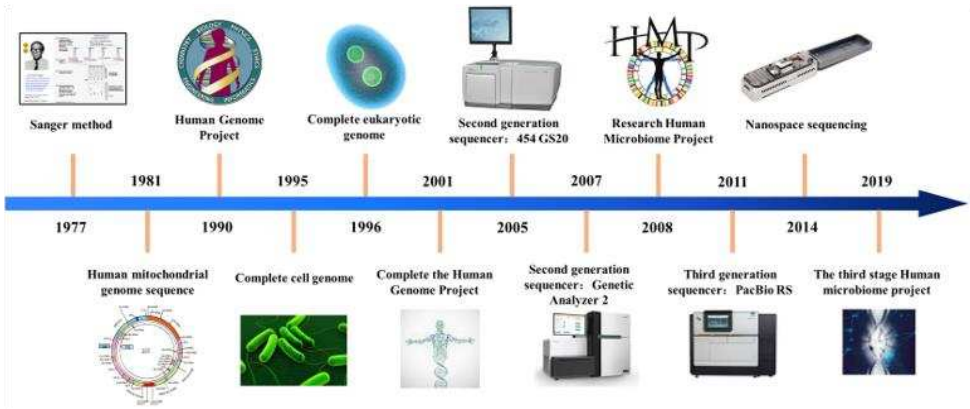


***Figure 3.2.1**. Historical development of sequencing technologies (Yang 2020)*

Over the past 40 years, significant advancements in sequencing technologies have had a major impact on DNA sequence variation research. As a result, the number and diversity of sequenced genomes have increased (*Figure 3.2.1*), while the cost per megabase went down (Mb) (*Goodwin 2016*).

The human genome project (HGP) initial draft was published in 2003 thanks to Sanger sequencing (*Figure 3.2.2*). After that, the first high-quality reference genome sequence was produced, making gene identification possible.

High-density microarrays are another technical advancement in the investigation of genetic diversity (*Figure 3.2.2*). Numerous short oligonucleotide probes target particular regions of the genome where SNPs that differentiate populations are located. As a result, it is possible to determine the genotypes of millions of variants with great precision and repeatability (*Nielsen 2011*). One of the most widely used arrays nowadays in population genetic studies is the Axiom™ Genome-Wide Human Origins 1 Array (*Keinan 2007)*. It covers ~630,000 SNPs, known to be variable in human populations, extracted from 13 population-specific panels, analysed in the Human Genome Diversity Project (HGDP) as well as the Neanderthal Genome Project (*Green 2010*).

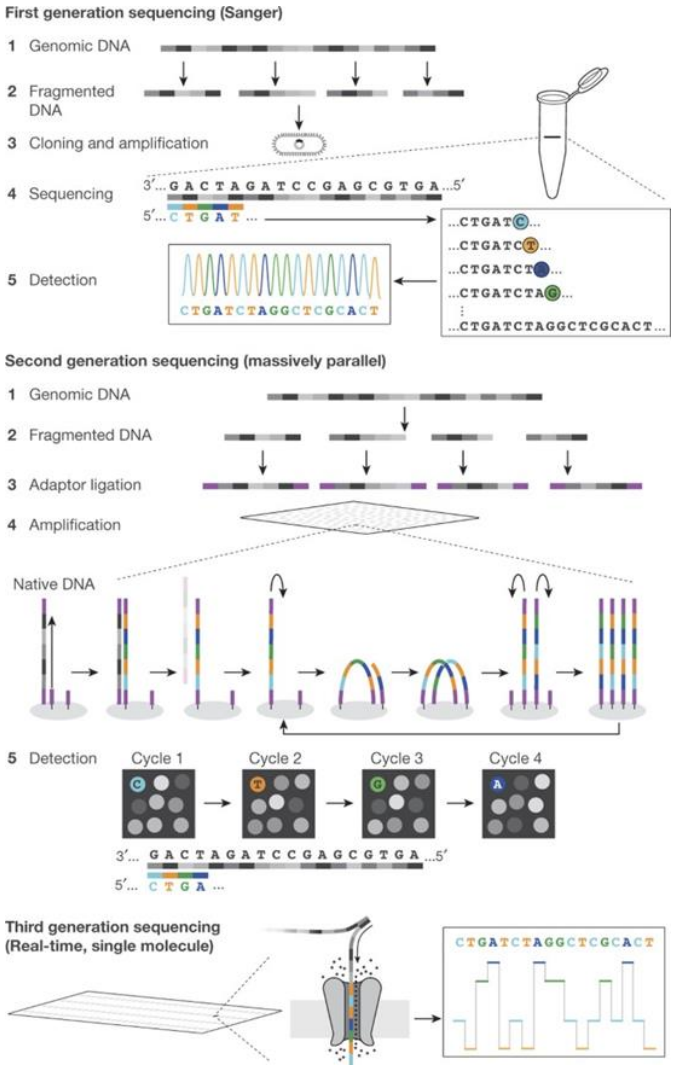


***Figure 3.2.2**. DNA sequencing technologies (Shendure 2017)*

A fundamental event for all genetics branches, which started a new era, was next-generation sequencing (NGS) (*Metzker 2010*), first commercially introduced in 2005 by 454 Pyrosequencing (*Margulies 2005*) and one year later Illumina Sequencing by Synthesis was presented (*Shendure 2008*). The second-generation technology (*Figure 3.2.1*) remains leading till nowadays thanks to its high throughput and accuracy, reducing the cost and the time needed for whole-genome sequencing (WGS). It gave a boost to such global projects as the 1000 Genome Project (1KGP) (*The 1000 Genomes Project Consortium 2015*), the Simon Genome Diversity Project (SGDP) (*Mallick 2016*), and the Human Genome Diversity Project (HGDP) (*Bergström 2020*). And from 2010 NGS technologies allowed the sequencing of whole ancient genomes (*Rasmussen 2010*).

Further development in technology after NGS (*Figure 3.2.1*), have led to the emergence of a new branch known as third-generation sequencing (TGS) (*van Dijk 2018, Shendure 2017*). TGS allows a direct sequencing of single DNA molecules without previous amplification. These technologies produce long reads, which are useful for resolving complex genomic regions, structural variants, and tandem repeats. Leading positions on third-generation sequencing technologies are held by PacBio (Pacific Biosciences) (*Eid 2009*) and Oxford Nanopore Technologies (ONT) (*Laver 2015*). It helped to fill gaps in the reference human genome (*Nurk 2021*).

3.3 Methods for analysing genomic data

Genomic data can be analysed by a variety of different methods, depending on the aim of the research. There are tools for sequence alignment, variant calling, *de novo* assembly, gene expression analysis or epigenetics, population genomics and many other fields. Three methods are commonly used to analyse genomic data: phylogenetic trees, principal component analysis (PCA), and admixture.

3.3.1 Phylogenetic trees

Phylogenetic trees represent diagrams with reconstruction of historical coalescence of individuals or populations through the analysis of variation that clusters samples into groups. These groups (taxa) derive from the common ancestor. The trees are composed of internal nodes and branches, which display the MRCA of each group and the genetic distance, or degree of kinship, between individuals or populations. The ability to infer genetic links between populations using uniparental markers, which do not take admixture events into account because of recombination, is a critical function of this technique (*Li 2011*).

Different methods can be used to build a phylogenetic tree. Distance matrix methods use a matrix with pairwise genetic distances between entities, such as genes or species. The distance shows the number of differences between each pair of entities. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA), which generates a unique phylogenetic tree, progressively combines the two taxa which have the lowest genetic distance between them. The Neighbor Joining (NJ) approach generates the minimalistic evolution tree, which represents the shortest sum of the branch lengths, which gradually creates branches until the last evolutionary relationship is found. The character state methods analyse the common and different traits among the entities: The Maximum Parsimony (MP) algorithm creates a tree with the fewest evolutionary changes; the Maximum Likelihood (ML) (*Figure 3.3.1*) assesses several tree topologies and, given an evolutionary model, determines the tree that has the highest likelihood of explaining the data (*Sharma 2018*).

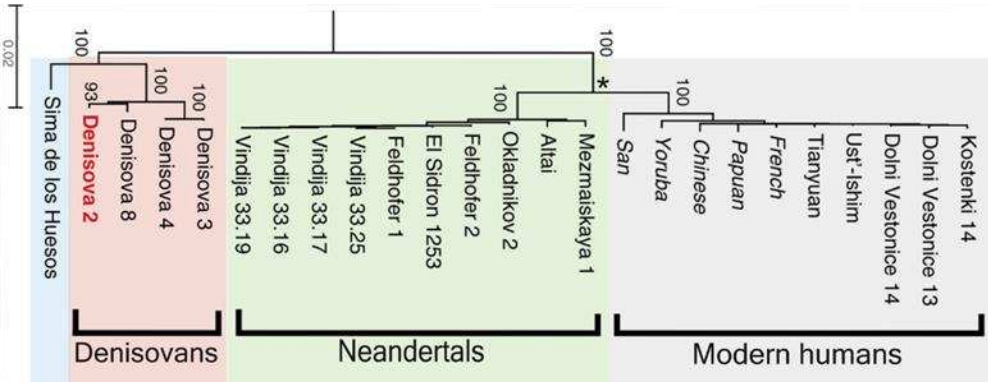


Figure 3.3.1.1. *Example of ML tree, modified from (Slon 2017). Phylogenetic tree relating the mtDNA from one Sima de los Huesos, four Denisovans, ten Neandertals and ten modern humans. The Denisova 2 mtDNA clusters with the three previously determined Denisovan mtDNAs, confirming its origin into the subspecies Denisova. The asterisk shows the MRCA of the groups Neandertals and modern humans*

Three more methods are particularly useful for studies of uniparental genetic systems (see paragraph 3.4). For mitochondrial DNA phylogeny, one method calculates the average distance between a group of sequences and their MRCA, assuming a constant mutation rate in all the branches of the tree. Another method, based on the ML approach, calculates the maximum likelihood of the DNA sequences based on the branch length, substitution and transition rates. The third one, a Bayesian method, estimates the highest likelihood of the tree hypothesis, using Markov chain Monte-Carlo (MCMC) simulations (*Soares 2009*).

3.3.2 Principal component analysis

Population structure can be measured and visualised using Principal component analysis method (PCA). This is a statistical technique which keeps most of the variability while simplifying the dimensionality of complex datasets (*Figure 3.3.2*). Calculations can be done in two ways: based on a correlation matrix or a covariance matrix derived from the genotype data (*Patterson 2006*). Principal components (PCs) capture the maximum variance in the data and are derived from eigenvectors (the directions in which the data varies the most), indicating the corresponding variance of each PC's eigenvalues (how much variance is captured in each of those directions). Usually, the first two strongest components account for most of the variation, which then defines location of the data on a coordinate system defined by this PCs.

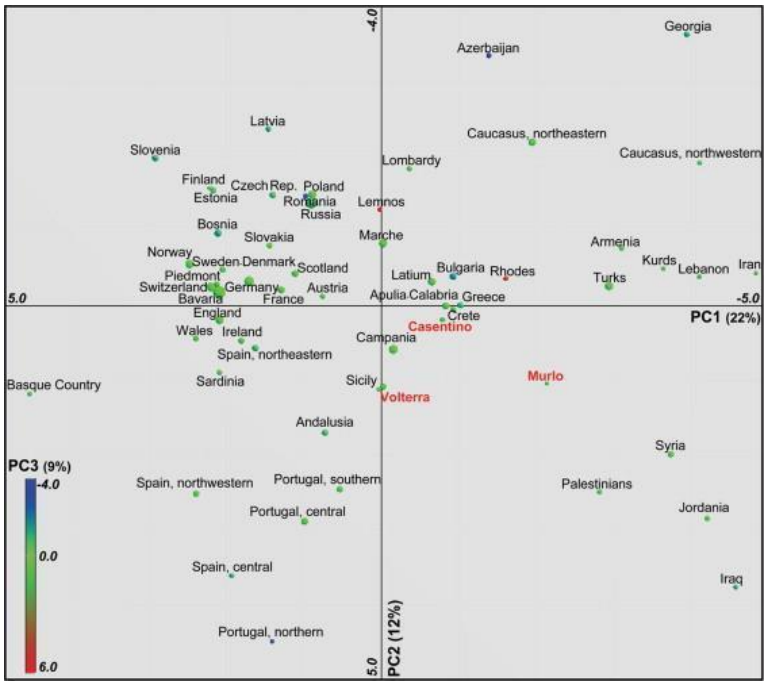


Figure 3.3.2.1. *Four-dimensional region-based PCA of mtDNA haplogroup profiles in Europe and the Near East (Achilli 2007)*

Therefore, PCA is a powerful tool for reducing data's complexity without sacrificing its essential structure, which facilitates analysis and interpretation.

3.3.3 Admixture

Another model-based clustering methods which are based on differences in allele frequency are ADMIXTURE (*Alexander 2009*) and STRUCTURE (*Pritchard 2000*). The algorithms are based on selectable K value, which defines the possible number of ancestral populations from which the population of interest derived (*Figure 3.3.3*). These techniques use separate or unlinked loci, typically needing a predetermined number of components (K), to which each genome is then allocated. It can be difficult to determine the ideal number of Ks, thus ADMIXTURE, for instance, uses a cross-validation technique to determine which K is best. More information on K value can also be obtained from EvalADMIX software (*Garcia-Erill 2020*) which estimates whether the selected model fits the described population ancestry.

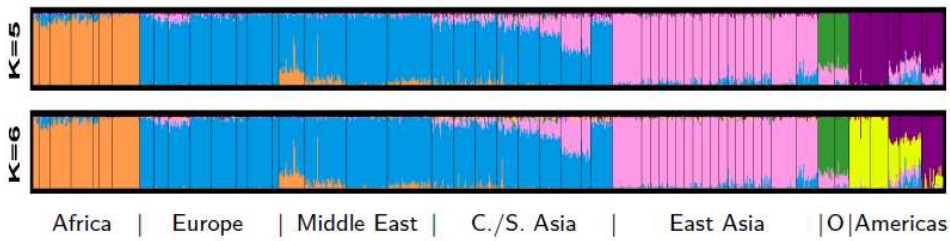


Figure 3.3.3.1. *Estimated population structure. The estimated membership fractions of each individual in each of the K clusters are illustrated by a thin vertical line that is divided into K coloured segments. Individuals from distinct populations are divided by black lines. Modified by Ida Moltke from (Rosenberg 2002)*

3.4 Uniparental markers

Despite the accessibility of a broad range of tools which gives ability to dive into the complete autosomal genomes and receive high resolution data, uniparental molecular systems remain very popular and highly informative sources of information in population genetics. These are genetic systems which are inherited from only one parent. In humans, uniparental molecular systems are represented by mitochondrial DNA (mtDNA) and by the Male Specific Region of the Y chromosome (MSY) (*Underhill 2007*). The mode of inheritance makes these two of particular value for studies of evolution, genealogy, medical genetics and forensics (*Henn 2010, Kivisild 2015, Wilkins 2006*).

Mitochondrial DNA can be inherited only through the maternal line (*Figure 3.4.1*), because, during fertilisation, of oocytes the mitochondria that

power the sperm are destroyed. Therefore, all children, despite their sex, inherit maternal mitochondrial DNAs (*Schon 2012*). The Y chromosome, instead, passes only from father to son, because this chromosome normally determines sex, and genes located on it, turn on the cascade of biochemical reactions which turns embryos into a male (*Skaletsky 2003*). The advantage of studying uniparental markers is the absence of recombination, which makes it possible to reconstruct the genealogies of single and/or groups of individuals, and to build phylogenetic trees, which, together with the phylogeographic approach, can be used to reconstruct the origins and migrations of populations (*Torroni 2006*). All changes that occur in the genetic material of mitochondria (that is, random mutations) accumulate and transmit through the maternal line. Over time, such mutations become fixed and remain stable in the gene pool, thus serving as reliable markers for classifying DNA into haplogroups. Each mitochondrial haplogroup (Hg) is a monophyletic unit that is defined by a set of variants inherited from the same maternal ancestor (*Gasparre 2020*).

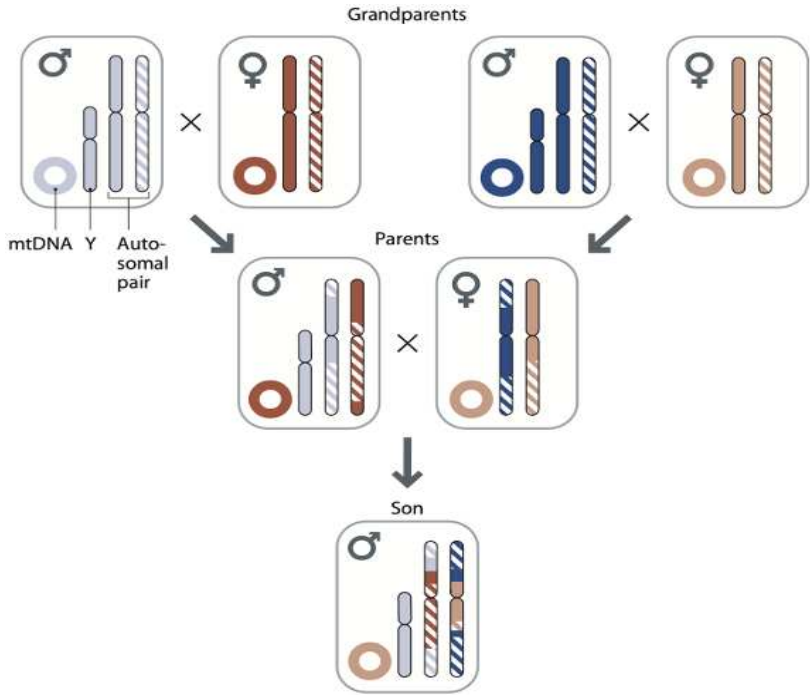


***Figure 3.4.1**. Scheme of inheritance of recombining and non-recombining segments of the genome (Jobling 2014)*

All variation presents nowadays in mtDNAs and MSY regions can be traced back in time to one ancestral molecule, which existed in some moment in the past (so called “mitochondrial Eve” and “Y-chromosomal Adam”) (*Poznik 2013*). The accumulation of new mutations in uniparental genetic systems is relatively fast (due to exposure to reactive oxygen in cellular

respiration for mtDNA; or lack of recombination, limited effective population size, founder effects and migrations for both MSY and mtDNA), and thus these mutations occurred and accumulated while modern *Homo sapiens* were colonising new lands and continents. Therefore, these mutations reveal remarkable traces of geographic and populational affiliations for different haplogroups, making them geo-referenced (*Torroni 2006*). Building phylogenetic trees, together with phylogeographical knowledge, is used to investigate the origins and migrations of populations, as well as their demographic histories (*Avise 2000*).

Since this PhD thesis contains a part where mtDNA is analysed, it is reasonable to give a deeper description on the mtDNA features and structure.

3.4.1 Genetic peculiarities of mitochondrial DNA

The human mitochondrial DNA molecule (16,569 base pairs, kb) is shorter than the nuclear genome (3.2 billion bases, Gb) and this feature simplifies the way of analysing, shortens the time of studying and reduces the costs of each research (*van Oven 2009*). Mitochondria are present in multiple copies in the cell and can contain several mtDNA molecules each, thus, every cell has many copies of mtDNA, from a few units to hundreds, depending on the cell type. Human mtDNA encodes 37 genes (*Figure 3.4.1*), including 13 protein-coding genes, two ribosomal RNAs, and 22 tRNAs (*Yan 2019*).

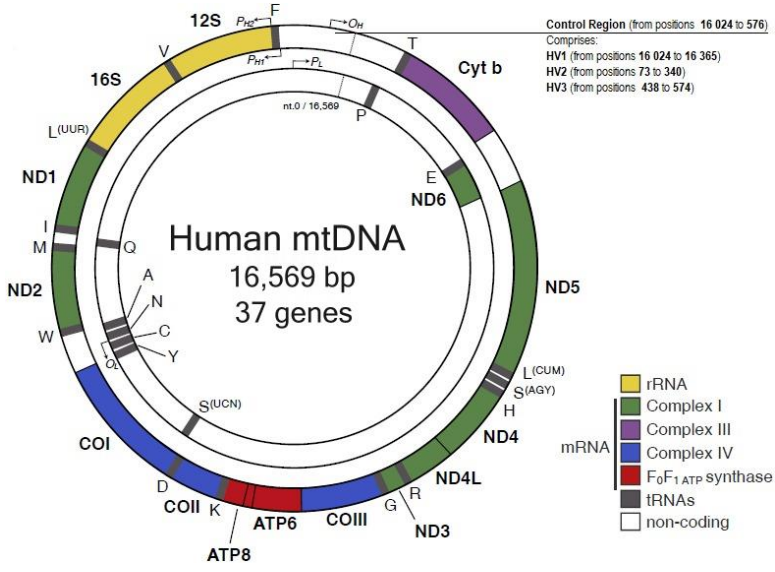


Figure 3.4.1.1. *The human mtDNA genome* (*Amorim 2019*)

The mtDNA molecule is circular and, according to the Endosymbiotic theory (*Gray 1999*), has a protobacteria origin. Therefore, it reminds prokaryotic nucleoid by the organization. The mtDNA molecule is double stranded and has a heavy (H) and light (L) strands, with differences in guanine and cytosine quantity. The H-strand is guanine-rich while the L-strand is cytosine-rich. Most of the genome is coding, while the primary non-coding portion is known as the "control region" or D-loop and it is responsible for controlling transcription and replication (*Brandstätter 2004*). As the most polymorphic region of the mtDNA, the D-loop spans 1,121 bps (nucleotide positions, nps, 16024–576) and contains three so-called hypervariable segments: HVS-I (nps 16024–16400), HVS-II (nps 44–340), and HVS-III (nps 438–576). Recombination does not occur in the mtDNA, as previously stated (*Hagström 2013, Hutchison 1974, Merriwether 1991*), but this molecule has a 10-20 times higher rate of mutation than that of nuclear DNA (*Brown 1979*). These features made mtDNA a crucial tool for evolutionary genetic studies for more than 40 years now.

The first sequenced human part of DNA was mtDNA, which became a reference sequence (Cambridge reference sequence, CRS), which was published in 1981. It was later re-sequenced and corrected for some nucleotides and ambiguities (*Andrews 1999*). Until now, this revised reference sequence (rCRS) is used as a standard reference for human mtDNA.

Restriction fragment length polymorphisms (RFLPs) served as the foundation for the mtDNA phylogeny from the 1990s, and alphabetic labels were used to identify the most prevalent haplogroups (*Torroni 1993*). For American and Asian branches were assigned Hgs from A to G (*Torroni 1993, Torroni 1994*), for Europeans from H to K (*Torroni 1994*[2]) and for African countries – L (*Chen 1995*). Up to date mtDNA phylogeny (http://www.phylotree.org/) coming from multiple studies on complete mtDNA sequencing across the world. The worldwide phylogeny on base of PhyloTree 17 had a recent update with 966 haplogroups to version named PhyloTree 17 Forensic Update 1.2 (*Dür 2021*).

Therefore, all these features of mtDNA and information background allows one to obtain knowledge about genomic variability only by sequencing a molecule much smaller than a chromosome. Moreover, human cells can contain up to 1000 mitochondria which, in turn, possess several copies of mtDNA each. Therefore, it can be analysed also when the biological specimen is poor on genomic DNA, i.e. archaeological finds and forensic samples (*Budowle 2003*).

3.4.2 Human mitochondrial DNA phylogeny

The mutation rate in mtDNA of vertebrate animals is not easy to estimate because mutation frequencies are not uniform and can be deteriorated by factors such as the presence of heteroplasmies (more than one type of mitochondrial DNA in a cell or entire organism) and genetic drifts, which lead to the loss of de novo mutations (*Rebolledo-Jaramillo 2014*). Genetic drift can also switch the direction of mutations as well as result in a higher effective transition/transversion rate biases (*Frank 1999, Bandelt 2006*). Nevertheless, according to two different estimates for the coalescence time of the mtDNA to the MRCA using more relaxed-clock models (paragraph 3.1), the age of ancestral mtDNA molecule (mitochondrial Eve) is estimated to be between 100 to 200 thousand years ago (kya) (*Behar 2012, Poznik 2013*).

The “mitochondrial Eve” has evolved in Africa, as African populations show the highest genetic diversity (the longest tree branches, *figure 3.4.2*). Sub-Saharan African populations belong to seven main haplogroups (L0-L6) (*Chen 1995*). Although, L3 haplogroup is also found in Mediterranean Europe and West Asia (*van Oven 2009*). Interesting fact, that L3 was the maternal hg for all non-African mtDNA lineages, affirms presence of a big bottleneck effect during the exit of humans “out of Africa” and their further migrations (*Cerezo 2012*).

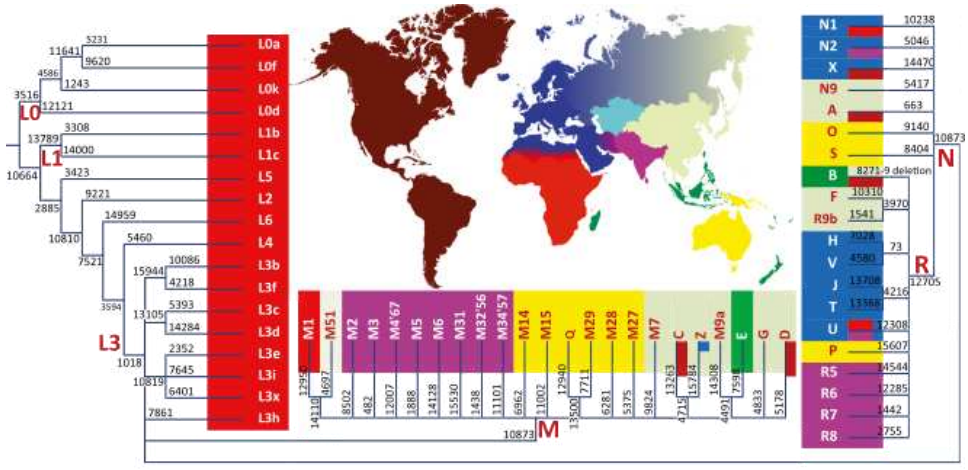


Figure 3.4.2. *Mitochondrial DNA tree and worldwide distribution map of the haplogroups from (Kivisild 2015). Labels are reported according to the nomenclature of PhyloTree. Geographic distribution corresponds to the haplogroup colours.*

First non-African haplogroups M and N separated from L3 approximately 62-95 kya (*Fu 2013*). These two clades diverged one from the other 40-70 kya (*Macaulay 2005*). They spread in Asia, Australia, Oceania, and Americas, while their sub-branches have more specific distribution. Remains with hg R were found in Western Siberia and the age estimate of this hg divergence is

dated at 45 kya (*Fu 2014*). In South Asia, particularly in India and in Southeast Asia exists the largest variety of sub-branches of hgs M, N and R (*Chaubey 2008, Hill 2007*). These findings suggested a possible way of leaving Africa by the South Asian route (*Quintana-Murci 1999*). The hgs U, HV, JT, N1, N2 and X are reported in Europe, Southwest Asia, and North Africa (*Soares 2010*). The hgs R5-R8 and M2-M6, A-G, Z and M7-M9 are found in East Asia and the Americas (*Stoneking 2010*).

Peopling of the Americas happened in the end of global migration of modern humans, relatively recently, and associated with at least three migration waves (*Reich 2012*). The first migration to the Americas dated back to 15-18 kya, and brought A2, B2, C1b, C1c, C1d*, C1d1, D1, D4h3a, and D4e1c lineages to both continents. In the second wave arrived people with C4c and X2a lineages to North America. And the third wave, approximately 5 kya, came D2a lineages to the Arctic through Northern Canada and Greenland, which were later replaced, by people with A2a, A2b, and D3 lineages (*Achilli 2013*).

Yet, all the research conducted in the past does not cover all the issues, such as the numbers and sources of Mesolithic and Neolithic gene flows in Europe, colonisations, peopling of Oceania etc. Nowadays all this and other open questions are being studied with not only uniparental genetic systems, but also on autosomal markers, especially with ancient DNA (aDNA), on macro- and micro-geographical levels.

3.5 Biparental markers

Genetic variation can also be studied by using higher resolution methods which concentrate on autosomal markers. While maternal or paternal lines can be interrupted, for example, if a father has only daughters, therefore he does not pass the Y chromosome to future generations, and his Y-line disappears with him. The same applies to mtDNA, which cannot be inherited further from sons to their offspring. Autosomes carry variants and information from all genetic ancestors of an individual.

The human genome is relatively large and complicated, it is organised into 23 pairs of molecules called chromosomes. The total size of the haploid nuclear genome is around 3.2 billion bp. The chromosomes are named by numbers, ordered by size, from 1 to 23, or from 1 to 22 and the last separated pair are the sex chromosomes (X and Y). Sex chromosomes are crucial for sex determination in humans, and according to the XY system, females have two X chromosomes, while males one X and one Y (*Graves 2006*). Autosomes and pseudoautosomal regions of sex chromosomes undergo meiotic recombination (crossover) (*Jeffreys 2001*), which allows to shuffle genetic

material every generation and creates new combinations for increasing variability of offspring. This mechanism works during the meiotic formation of gametes, which is going through two particular cell divisions, reducing the full set of chromosomes by half. Thereby, each gamete, under normal conditions, carries only a haploid set of chromosomes, which gets restored during the fertilisation process (*Zickler 2015*).

Complications in studies on autosomes come not only from crossover, but also from the presence of different types of genes, which can either be linked or not linked, and can always be inherited separately, even if they are located on the same chromosome. These gene combinations can be inherited from the same parent as haplotypes. Usually, the degree of linkage can be described by the distance between genes, however, in some regions of chromosomes, the recombination frequency is much higher than in the others, and these regions are called hotspots. Highly recombining regions are small (1-2 Kb) and are usually separated by larger segments with low recombination activity (*Jeffreys 2001*).

All these details must be considered when performing population genetic analyses of complete nuclear genome or genome-wide SNP chips.

3.6 The ancient DNA revolution

The genomes of present-day individuals can already give us quite many answers about past events, hypothesising how genetic information has changed based on reconstructing models. However, the advent of the ancient DNA field, which began a few decades ago, has opened windows directly into the evolutionary events of the past (*Tuross 2018*).

Ancient DNA can be defined as DNA extracted and analysed from fossils and archaeological remains (such as bones, mummies, etc.). The first attempt of sequencing of aDNA was made in 1984 from the extinct species of horse family, a *quagga*. This fragment was very short, just 229 bp (*Higuchi 1984*). Sequencing of complete ancient genomes became available only after the development of NGS technology (*Metzker 2010*). Thus, the archaeogenomic revolution (*Achilli 2018*) started in 2010 when the first human (*Rasmussen 2010*), Neanderthal (*Green 2010*) and Denisovan (*Reich 2010*) genomes were sequenced. Since that time tens of thousands of human and other organisms' ancient genomes have been sequenced and published (*Bergström 2020, Daly 2021, Frantz 2020, Kocher 2021, Librado 2021, Spyrou 2019, van der Valk 2021, Reich 2010*).

Working with aDNA is challenging due to various degradation processes, such as physical fragmentation, post-mortem DNA damage, and modern

DNA contamination. One of the most frequent chemical hydrolytic modifications of aDNA is the deamination of cytosine. This nucleotide becomes a uracil, when losing its amino group (-NH2). Cytosine is a pyrimidine which is losing the amino group in a faster and easier way compared to other types of nucleobases. Such substitutions happen more frequently at the ends of fragmented DNA molecules where they are more exposed and vulnerable to chemical changes. Consequently, while replicating aDNA fragments, according to complementary base-pairing rules, uracil will be paired with adenine, resulting in thymine appearance on the complementary strand (C to T nucleotide substitutions) (*Figure 3.6.1*). As a result, GC pair is substituted with AT, accumulating these "phantom" mutations, which must be taken in consideration while analysing aDNA (*Hofreiter 2001*).

These changes in DNA show a correlation with time and environmental surroundings. If specimens were found in the same environmental conditions but they belong to different ages, the older specimen will show more post-mortem damage. When the environment is more humid and warmer the changes in DNA structure are much faster than in cold and dry places (*Sawyer 2012, Wagner 2018, Weiß 2016*). Nevertheless, DNA damage can bring not only problems for the geneticists, but also it has become an important instrument for distinguishing endogenous DNA from modern contaminants (*Pääbo 2004*).

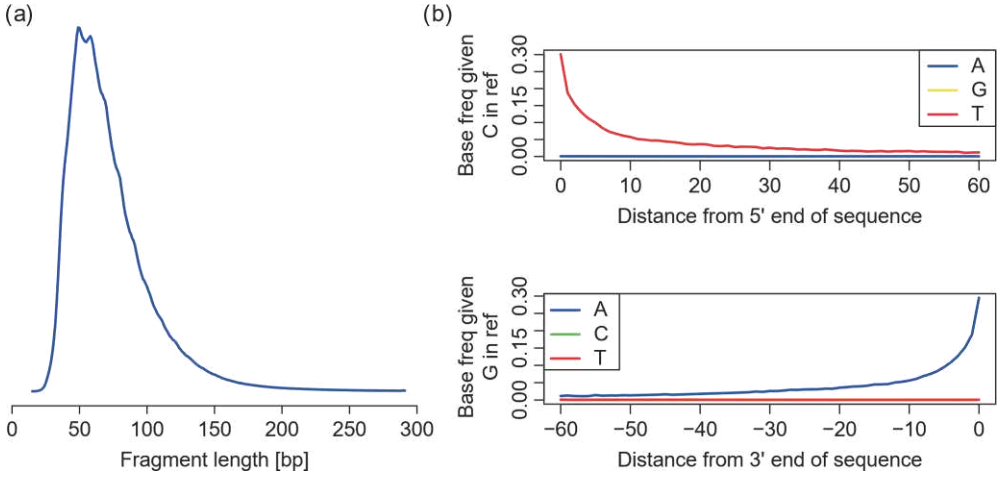


Figure 3.6.1. *Characteristic damage pattern in aDNA: (a) length distribution of sequenced fragments; (b) enrichment of deamination toward the fragments ends (Balding 2019).*

Excavation, further transportation of remains, and laboratory analyses all pose risks for contaminating ancient human samples with modern human DNA. In addition, contamination from soil bacteria or other microorganisms that accumulate after the death of the organism, is far easier to detect and

handle. The concentration and quality of these contaminant DNAs are usually much higher than for the ancient DNA, presenting a significant challenge when working with it. If the studied species belongs to other animal branches, distant from modern humans, it is easier to remove contamination by mapping the sequenced genome to a specific reference and detecting the contamination sources by mapping against potential contaminants. Therefore, the aDNA damage pattern has become a handful method for detecting authentic ancient human DNA (*Renaud 2015*).

Modern methods of analyses triggered by NGS and development of bioinformatics opened the age of aDNA investigations. Such meaningful breakthroughs like sequencing of a one-million-year-old mammoth (*van der Valk 2021*), and the sequencing of distinct hominin species and modern *Homo Sapiens* dating back as far as 45 kya have improved the knowledge about our past. This data allows us to calibrate molecular clocks for different phylogenies and make them more precise (*Fu 2014, Helgason 2015, Posth 2016*).

It is worth mentioning that the further development of methods for working with such complicated DNA samples, like ancient DNA, allows nowadays to retrieve DNA from samples with very low endogenous content (even below 1%). For these, specific capturing techniques and laboratory equipment must be used. This can be achieved with enrichment methods such as deep sequencing, single-stranded DNA library preparation, hybridization capture, and selective amplification. These methods selectively pull out ancient DNA from a contaminated background by improving the signal-to-noise ratio (*Carpenter 2013, Maricic 2010*).

3.7 The complex history of present-day Ukraine

The migrations of *Homo sapiens* around the continents took place unevenly, and these processes took place in a wave-like manner, with different intensity and at different intervals of time. The peopling of Europe and the population diversity of the continent is one of the most studied topics in population genetics *(Haak 2015, Allentof 2015, Mathieson 2018, Hofmanová 2016, Omrak 2016, Olalde 2018, Lazaridis 2018)*. Anatomically modern humans left Sub-Saharan Africa about 60 kya. At that time admixture with Neanderthals had occurred. Migrations brought the first known settlers to Europe about 45 kya, according to archaeological evidence. However, recent findings suggest even earlier settlements of anatomically modern humans in Europe, particularly in France, around 55 kya (*Slimak 2022*). In fact, these first settlers did not contribute to the formation of the genetic pool of modern Europeans. Later, a younger settlement was found with remains related to present-day Europeans (Kostenki 14 – 37 kya; Buran-Kaya - 36-37 kya; Goyet Q116-1,

Bacho Kiro 1653 – 35 kya) (*Lazaridis 2018, Bennet 2023; Posth 2023*). Different cultures emerged across Europe, with the Last Glacial Maximum (LGM), around 26 to 19 kya, having a significant impact on the genetic makeup of populations. Many populations of that era were replaced or admixed with the wave of hunter-gatherers with strong components from the Near East (Villaburna cluster) (*Lazaridis 2016*). Later, around 8-9 kya, a period of big cultural and behavioural changes, marked as "Neolithic revolution", occurred. European populations experienced a significant influx of new migrants from the Near East, who were mainly represented by farmers. During this period many animals and plants were domesticated. (Oberkassel cluster) (*Aneli 2021*).

From Eneolithic time, the Trypillia culture flourished in the central and western regions of present-day Ukraine, from approximately 7-5 kya. This culture is known for its exceptionally large settlements, sometimes referred to as proto-cities, some of which housed thousands of inhabitants, a remarkable size for the time. The Trypillia culture influenced neighbouring groups and eventually declined for reasons that remain uncertain, possibly due to environmental changes, social factors, or interactions with neighbouring nomadic groups (*Schmidt 2020*). Importantly, the Trypillia cultural complex played a pivotal role in the spread of farming into Eastern Europe (*Gelabert 2022*).

During the early Bronze Age, another significant migration event occurs (*Haak 2015*). Populations from the Caucasus reached Eastern Europe and, together with local populations, formed steppe groups. Around 4.9 kya, the Yamna culture emerged in the Pontic-Caspian steppe and expanded both westward and eastward, which gave rise to the Corded Ware culture in central Europe (*Haak 2015*). This culture coexisted for some time with another, equally important, Catacomb culture. Since that period, the formation and admixture of various cultures have continued, contributing to the present-day European gene pool. Modern Europeans carry a substantial genetic influx from Early Bronze Age steppe pastoralists, Neolithic farmers from the Near East, and Upper Paleolithic hunter-gatherers (*Richards 2016*).

The area of present-day Ukraine has been a crossroads of people and cultures since Palaeolithic times because of its strategic location on the northern shore of the Black Sea at the intersection between East Europe and the Pontic Caspian steppe. The territory of Ukraine was also a cradle of civilization for Slavs and Crimean Tatars, both of which massively contributed to peopling of the countries located in East Europe (*Abdulaeva 2021, Yakubova 2014*). In more recent times, its strategic location allowed it to develop markets, and ports. Therefore, the past population(s) of Ukraine experienced a high number of admixture events, which make it of great interest to untangle the knot that wraps the origin and genetic history of Ukrainians in a lot of different layers. Moreover, these features make the

assessment of Ukrainians' genetic variation of particular value to fully reconstruct the genetic landscape of Western Eurasia.

Since the ancient individuals studied in this thesis belong to three major ancient cultural groups, a brief description of each is provided below.

The Yamna (Pit-grave) archaeological culture appears around six kya, named after its particular burial tradition. Burials typically contained a single individual in a supine position with bent knees. On the remains could be found red ochre, and they could be accompanied with different grave goods such as tools, weapons, and ornaments, reflecting social differentiation. Tombs were often covered by a kurgan. Geographically, the Yamna culture was spread across parts of modern-day Ukraine**,** southern Russia**,** Moldova and eastern Romania. Later, it expanded westward into the Balkans and eastward to the Volga region, interacting with other cultures. In the North Pontic area, this period is roughly spanned from 3000 to 2300 BCE (from the Early Bronze Age (EBA) to the Late Copper Age (LCA)) (*Rassamakin 2008, Rassamakin 1999, Telegin 1999*), while in Caspian-Ural region it was recorded even earlier. The Yamna people worked with bronze and copper to produce tools and weapons, though they also used stone tools, flint daggers, copper and bone ornaments. The tribes were nomadic or semi-nomadic, moving across the steppe with their cattle, sheep, and goats. Their mobility was enhanced by horse domestication and use of horse-drawn chariots. The Yamna culture is also associated with the Proto-Indo-European language, which is ancestral to many modern South Asian and European languages (*Anthony 2007*).

The Catacomb culture occupied a territory roughly similar to that of the Yamna, extending into Lower Don and Lower Volga regions. It emerged later than Yamna, but for some short period of time both cultures coexisted. The Catacomb culture also took its name from a particular burial practice which featured a side chamber under a primary gravel pit. The body position, grave goods and presence of kurgans are similar to those found in the Yamna culture. Metalwork on bronze and copper indicates that these materials were essential to their economy and daily life. In the North Pontic steppe, it is dated back to 2450 to 2000 BCE (*Telegin 1999*). Early monuments of the Catacomb culture in the Northern Black Sea region are approximated to date from 2800 to 2500 BCE (*Pustovalov 1994*), while the Late Culture monuments are dated to 2450-2400 BCE (*Govedarica 2006*) with extreme dates ranging from 2000 to 1900 BCE (*Shishlina 2009*). The Catacomb culture was primarily pastoral with some evidence of limited agriculture and its people engaged in hunting, fishing and gathering. Social stratification is indicated by burials containing specific grave goods for warriors and leaders. Additionally, fortified settlements with defences suggest that conflicts between different groups were a part of their daily life. This culture played a significant role in the transition from EBA societies to more complex in later periods (*Anthony 2007*).

The Scythian culture was represented by ancient nomadic civilization that thrived in the Eurasian steppes. It emerged in the North Pontic-Caspian region in the first half of the 7th century BCE, which is regarded as the core zone of the Scythians. Their territory encompassed modern-day Ukraine, including Crimea, southern Russia and part of Kazakhstan, while their broader sphere extended to the lands of Anatolia and regions along the Danube river. The Scythians were mostly nomadic pastoralists herding large numbers of cattle, horses and sheep. They were exceptionally talented riders and were recognized for the invention of several aspects in mounted combat, including cutting-edge cavalry strategies. Ancient sources frequently depict the Scythians as fierce warriors capable of launching fast and lethal raids (*Sulimirski 1995*).

Their military strength allowed them to confront powerful empires such as the Persians and Greeks, impose tribute on surrounding nations, and exercise authority over enormous swaths of land. A common trait among the Scythians and other related cultures was the tradition of kurgan burials which often featured richly furnished graves with different items. Notably, elite Scythian burials included horses that were sacrificed and buried alongside the owners. Interesting fact, that the burials frequently contained artefacts made of gold, which can reflect the craftsmanship and their connections to trade networks. The Scythians played a crucial role in cultural exchanges between the East and the West, influencing and being influenced by the art, technology, and trade patterns of other ancient civilizations. However, the decline of the Scythian culture started in the 3rd century BCE, as intense movements of new cultural groups, such as Sarmatians and other tribes in Central Asia, began to displace them in most parts of the territory (*Gerling 2015*). By the early 2nd century BCE, only a small Scythian population remained in the Crimea (*Parzinger 2009*).

3.7.1 Focusing on Donetsk oblast

Donetsk oblast (region) is located in the Great Eurasian Steppe, on the border between the nomadic pastoralists and more sedentary communities. This geographic position contributed to the mixing of the populations with diverse gene pools. Recent migration processes in this region were particularly dynamic during the 19th and 20th centuries, driven by urbanisation and social modernization under the occupations of the Russian Empire, the Soviet Union, and finally, independent Ukraine (*Yakubova 2014*). Nowadays it borders the Russian Federation, and still a big part of the region has been under Russian occupation since 2014. This context makes it particularly important to reveal the events of migration and gene flows with surrounding regions that left their footprint in the formation of the local population.

Donetsk oblast is a large region located in southeastern Ukraine, covering 26.517 km^2 (4.4% of the total Ukrainian territory). This province is the most

populated of the entire country, with a population approximately 4.1 million inhabitants (*State Statistics Service of Ukraine 2001*). Geographically, Donetsk oblast is located in the Eurasian steppe landscape, surrounded by woods, hills, spoil tips, lakes, rivers; with access to the Sea of Azov to the south *(www.encyclopediaofukraine.com).* The region of modern Donetsk oblast was actively settled by Ukrainian Cossacks in the 16th century, with many villages fortified by the Zaporizhian Cossacks to defend against the Crimean Tatars and the Ottoman Empire (*Plokhy 2015*). Later, from the 18th century, this land was annexed by Russia which initiated significant restructuring and policies aimed at the colonisation of the native population.

The main city of modern Donetsk oblast is Donetsk, founded in 1869, by John Hughes (a Welsh businessman) who established a steel plant and owned several coal mines in the region. First name of the settlement was Yuzovka, from Slavic simplification of Hughes's surname, "Yuz". The city rapidly grew due to industrialization, particularly around coal mining and steel production, during the 19th and 20th centuries. In 1961, the city was renamed to Donetsk and became the administrative centre of Donetsk oblast (*https://www.britannica.com/place/Donetsk-Ukraine*). Donetsk oblast became a key part of the USSR's heavy industry. To break the Ukrainian resistance to Soviet authority and accelerate politics of collectivization, many areas of Eastern Ukraine, including Donetsk region experienced a man-made famine, known as the Holodomor in 1932-1933 (*Applebaum 2017, Plokhy 2015*). During World War II, the region suffered significant destruction and loss of life. As a result of these events, the indigenous population was rapidly replaced by migrants from various regions of Russia (*Applebaum 2017, Yakubova 2014*). Furthermore, the socio-political situation over the last ten years has led to substantial changes in the population composition and economic conditions, impacting the region's development. Thus, the collection of samples conducted in 1998-1999, used for this thesis, provides unique and valuable evidence of the various historical events that have shaped this population.

3.8 Population genetics studies on Ukraine

The study of mtDNA variation in different populations around the world, including those in Europe, has intensified over the past few decades, though it remains far from complete. Currently, further advancements in mtDNA phylogeny can be primarily achieved by investigating small populations and specific regions at a micro-geographic level (*Cardinali 2022*).

According to previous mtDNA studies (*Pshenichnov 2013, Malyarchuk 2001, Malyarchuk 2023*), the main haplogroups found in Ukraine were typical of Western Eurasia, including H, V and other sub-lineages of HV, J, T, I, N1b,

W, and X. Additionally, various sub-branches of U, i.e. U4, U5a (U5a1 and U5a1a) typical of the Volga region were observed. These data point out similarities between the Ukrainian gene pool and other European populations. However, haplogroups of East Asian origin, such as A, B, C, D and G, were also found, typically between one to three percent of the studied cohort.

It should be noted that previous studies on mtDNA variation in Ukrainians primarily were focused on the analysis of nucleotide sequences of HVS1 and HVS2, as well as the distribution of mtDNA haplogroup frequencies in populations *(Malyarchuk 2001, Nikitin 2009, Балановский 2012, Mielnik-Sikorska 2013, Pshenichnov 2013).* Georeferenced analyses of mtDNA polymorphism have demonstrated that East Slavic populations show a significant degree of affinity among themselves. Specifically, the gene pool of Ukrainians is most similar to those of populations in Poland, Belarus, and southern Russia (*Балановский 2012, Pshenichnov 2013).* Additionally, studies have highlighted that certain western Ukrainian populations living in mountainous regions exhibit greater genetic similarity to western Slavs than to other Ukrainians (*Nikitin 2009, Kushniarevich 2015*).

Therefore, the investigation of the mtDNA diversity based on data on the variability of entire mitogenomes from the population of Donetsk oblast is crucial for exploring the genetic history of this eastern region of Ukraine. This research will help to refine its genetic history and to better understand the dynamics of population movement and demography in the area.

4. Aims

Present-day Ukraine has been a crossroads of peoples and cultures since the Paleolithic, due to its strategic location on the northern shore of the Black Sea, between Eastern Europe and the Pontic-Caspian steppe. It served as one of the glacial refuges (the Eastern European Plain) where populations survived during the Last Glacial Maximum and later repopulated Europe. Ukraine was also the cradle of the Slavic and Crimean Tatar civilizations, both of which contributed significantly to the gene pool of Eastern Europe. Its strategic location, fertile lands, and favourable climate attracted migrations, settlements, and conflicts. These have resulted in a high number of admixture events that have layered the genetic history of the Ukrainians, making it valuable to untangle this complex genetic variation for reconstructing the genetic landscape of Western Eurasia.

Nowadays, the Donetsk oblast, located in eastern Ukraine, borders the Russian Federation, and much of the region has been occupied by Russian forces since 2014. In the past, its location on the edge of the Great Eurasian Steppe attracted both nomadic pastoralists and more sedentary communities. In the 19th and 20th centuries, the same area also experienced intense migration processes triggered by urbanisation and social modernization. Considering this history, it is clear why there is interest in disentangling the migration events and gene flows that left their own footprint in the formation of the local population. Therefore, one of the aims was to investigate the genetic variability of the modern population of Donetsk oblast (in the year 1999) for reconstructing its origins and history, by studying the complete mtDNA sequences and, for selected samples, genome-wide data to achieve higher molecular resolution and statistical power.

A second goal of this thesis is to provide a diachronic picture of genomic variation in the territory of present-day Ukraine at different times in the past. This is made possible by the availability of ancient remains related to three different cultures: Yamna, Catacomb, and Scythian. The analysis of ancient DNA obtained from these individuals will provide sequential genomic snapshots that will help to uncover the ancestral origins of the first settlers and subsequent arrivals, and to evaluate the effects of various gene flows and admixtures that occurred over time, ultimately adding the missing piece of genomic information on the Ukrainian population to the complex genetic landscape of Eastern Europe.

5. Materials and Methods

5.1 Modern Individuals

We received genealogical data for 104 male individuals, along with written informed consent for genomic studies. This collection was provided by our collaborators with Prof. Svetlana Arbuzova at the International Medico-Genetic Centre, Hospital Nol, Donetsk and Prof. Ornella Semino at the University of Pavia. A total of 104 biological samples from male individuals (aged between 18 and 75 years) were collected in Donetsk oblast (Ukraine) in 1998-1999, together with pedigree charts. Subjects participated voluntarily and provided written informed consent for genomic studies. Successfully extracted DNAs were preserved at - 20 °C in plastic tubes. Historical parts of the study were discussed with Yevhen Zakharhcenko (PhD, senior lecturer), Department of History of Ukraine of V. N. Karazin Kharkiv National University.

Nowadays, Ukraine's population is composed of more than 100 national or ethnic groups (*Kubicek 2008*). However, the main part of the population was represented by ethnic Ukrainians (78%) and ethnic Russians (17%) in 2001 (*http://2001.ukrcensus.gov.ua/results/general/nationality/*).

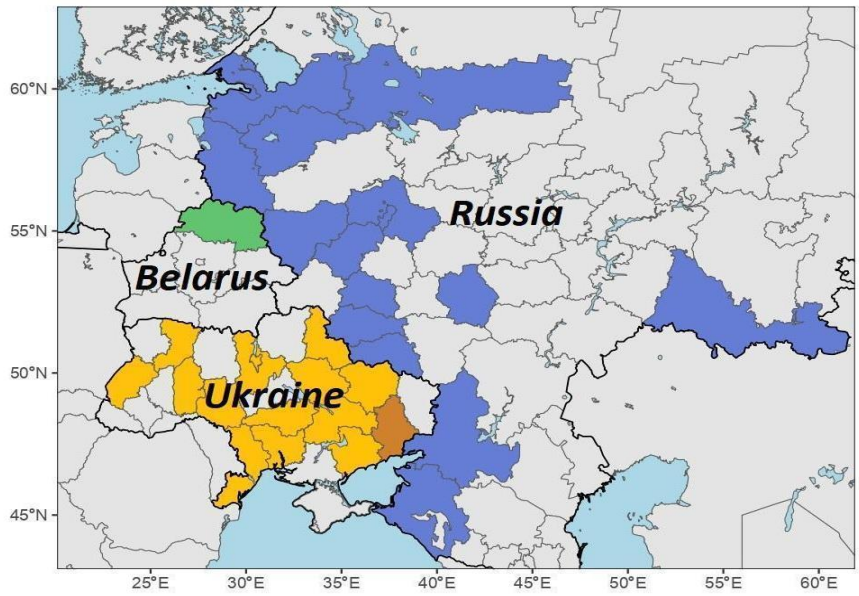


Figure 5.1.1. *Geographic distribution of the last known terminal maternal ancestors by regions (n= 104). (Donetsk oblast, Ukraine, the sampling region. Other colours show regions of origin of TMAs according to the country)*

The 104 individuals showed different TMAs, originating not only from different regions of Ukraine but also from two neighbouring countries, Russia and Belarus (*Figure 5.1.1*).

Among the 70 individuals with a Ukrainian TMA, 33 people marked Donetsk oblast as the birthplace of their grandmother, six had TMAs from both Zaporizhzhia and Vinnytsia oblasts, four from Kyiv and Kirovohrad oblasts each, three from Dnipropetrovsk and Kharkiv oblasts each, and two each from Sumy, Odesa, and Poltava oblasts. Mykolaiv, Khmelnytskyi, Rivne, Lviv, and Chernihiv oblasts were represented by one person.

Russian TMAs were registered for 33 individuals, and only one had a TMA from Belarus. Some regions of Russia are not shown on the map due to their great geographical distance: Omsk oblast, Amur oblast, and Irkutsk oblast. Other Russian regions were represented in the following way: six people each from Kursk and Rostov oblasts, three from Oryol and Smolensk Oblasts each, and two each from Tambov and Pskov oblasts. The remaining Russian regions, including Amur oblast, Belgorod oblast, Irkutsk oblast, Kaluga oblast, Krasnodar krai, Leningrad oblast, Moscow oblast, Nizhny Novgorod oblast, Omsk oblast, Orenburg oblast, and Volgograd oblast, were each represented by only one individual.

All individuals which were analysed in this study were redistributed according to the birthplace of the last known terminal maternal ancestor (TMA), as reported in (*Figure 6.1.1*, Chapter 6.1). It is worth to mention, that one third of probands who reported genealogical information had a TMA from Russia. This is in agreement with the 2001 census that estimates that Russians make up 38% of present-day Donetsk oblast population due to relocation of people during the existence of the USSR and after its fall (State Committee of Statistics of Ukraine). This agrees with historical data, since these Russian regions were very poor, thus pushing people to migrate to industrially developed places seeking for a job.

The provided data were analysed not only according to TMA origin, but also for the mitochondrial DNA complete genome study, while the Human Origin array sampling also considered the origin of terminal paternal ancestors (TPA).

5.1.1 Modern mitochondrial DNA

A total of 94 DNA samples, out of 104, were still available for this thesis. These DNAs had been extracted in a previous project studying the Y chromosome (*Semino 2000*), using the classic phenol-chloroform protocol with Corex centrifuge tube and TE 10:1 buffer. The extracts were quantified using the Invitrogen Qubit™ fluorometer.

5.1.1.1 Amplification and sequencing

The complete mtDNA was amplified in two partially overlapping fragments of ~8-9 Kb using Long-range (LR) PCR amplification. Primers for first amplicon were 5,193 Forward (sequence: CCCTTATTCCATCCACCCT), 13,829 Reverse (sequence: AGTCCTAGGAAAGTGACAGCGA), and for the second amplicon they were 13,477 Forward (sequence: AGGAATACCTTTCCTCACAGGTT) and 6,151 Reverse (sequence: GTGGTAAGGGCGATGAGTGT). To carry out this reaction, the Promega GoTaq® Long PCR Master Mix kit was used, with a reaction mixture that contains, in addition to the cofactors and buffers necessary for the correct functionality of the reaction, a thermostable and highly performing *Taq* polymerase. The final volume for each sample, including DNA, is 25 µl (Table 5.1.1.1).

Table 5.1.1.1. *Summary scheme of the initial and final concentrations and volumes for each reagent, used in the whole mitogenome amplification reaction.*

Reagents	Initial conc.	Volume (µl)	Final conc.
GoTaq® Master Mix	2X	12.5	1X
Primer For	10 µM	0.5	0.2 µM
Primer Rev	10 µM	0.5	0.2 µM
H2O		11.5	
Total		25.00	

Table 5.1.1.2. *PCR protocol used for Long Range amplification of DNA samples.*

Step	Temperature (°C)	Time	Cycles
Initial denaturation	94	2 min	1
Denaturation	94	30 sec	20
Annealing	58	30 sec	
Elongation	65	10 min	
Denaturation	94	30 sec	10
Annealing	55	30 sec	
Elongation	65	10 min	
Final extension	72	10 min	1

The amount of DNA required was evaluated following the quantification of the initial extract with Qubit 4™ Fluorometer; the volume of DNA therefore varied between 1 and 5 µl, keeping the final volume constant. The used PCR program has two consecutive reaction cycles, to allow the correct amplification of both long-range fragments (Table 5.1.1.2).

An agarose gel electrophoretic run was performed to verify correct amplification. A ladder of Sharpmass™ 1 kb plus ladder (*Euro Clone*) was loaded to control the amplicon size; this can also be used for the reading of Long-range fragments, as it has bands in a range between 10000 and 250 bp. The percentage of agarose used was 1%, since large fragments (8-9 kb) were analysed. After this evaluation step, LR PCR products were used to prepare NGS libraries, using the Illumina Nextera® XT DNA library preparation kit. The quality of prepared libraries was checked on an automated capillary electrophoretic platform Agilent Tapestation 4150, and then sequenced on an Illumina NextSeq 550 platform at the Genomic and Post-Genomic Unit, IRCCS Mondino Foundation in Pavia, in collaboration with Dr. Stella Gagliardi and Dr. Rosalinda Di Gerlando, as in previous works (*Brandini 2018, Modi 2020*).

5.1.1.2 Sequence analysis

Raw data were analysed using mtdna-server.uibk.ac.at (Weissensteiner 2016) and in-house scripts specifically developed by Dr. Nicola Rambaldi Migliore (https://github.com/NicRamb/mtDNA_pipe) to obtain the final haplotypes versus the revised Cambridge Reference Sequence (rCRS) (*Andrews 1999*). Ambiguous regions were manually checked. Three samples have been excluded due to the bad quality of sequences. Haplogroup classification and a maximum parsimony tree were obtained with HaploGrep v3.2.1 (*Schönherr 2023*), a classification tool which is based on the latest worldwide mtDNA phylogeny (*PhyloTree 17 Forensic Update (FU) 1.2*). The .fasta files were produced with HaploGrep v3.2.1 and used for downstream analyses.

5.1.1.3 Phylogenies

Phylogenetic trees for mitochondrial DNA were built using a maximum parsimony approach with the mtPhyl software (*eltsov.org/mtphyl.aspx*). The classification used in the mtPhyl is based on the PhyloTree 17 worldwide phylogeny (*Van Oven 2009*). These trees were manually corrected to reflect recent improvements in the mtDNA phylogeny (*Dür 2021*). Trees were also built with a Bayesian inference approach by using the software BEAST2 (Bayesian Evolutionary Analysis by Sampling Trees v.2.7.7) (*Bouckaert 2019*). This phylogeny was built with the HKY substitution model and a fixed molecular clock. Mutation rate was considered as in (*Fu 2013*). Coalescent age estimates were also calculated through BEAST2 using their

radiocarbon dates of mtDNA haplogroup L3a as priors. Subsequently, the resulting trees were processed with TreeAnnotator to obtain a merged tree. The tree was visualised using the FigTree software (*http://tree.bio.ed.ac.uk/software/figtree/*). Additionally, BEAST2 provides a demographic trend output, which displays the variation in effective population size (N_e) over time. Tracer v1.7.2 was used to visualise this trend as a Bayesian Skyline Plot (BSP).

5.1.2 Genome-wide analyses

5.1.2.1 The Human Origins (HO) array

The DNA of 45 out of 94 modern Ukrainian individuals from Donetsk oblast were selected according to the origin of their grandparents (all four from Ukraine) for genotyping with the Axiom Genome-Wide Human Origins 1 Array (HO) at the Institute of Health Research institute of Santiago de Compostela (CEGEN). It provides a genotyping panel specifically designed for population genetics, with genomic markers from 11 different modern human populations and ancient hominin individuals. Approximately 629,000 SNPs can be typed according to the samples of known ancestry, including data from Neanderthals, Denisovans and chimpanzees, avoiding the ascertainment biases which can be observed in genotyping with GWAS arrays. (*https://assets.thermofisher.com/TFS-Assets/LSG/brochures/axiom_hu origins_datasheet.pdf*).

All 45 genotyped individuals passed the quality checks with a call rate higher than 97%. The Axiom Analysis Suite V5.3.0.45 software was used to analyse the raw genotyping data. The resulting variant calling format file (VCF) was converted to plink format using PLINK v.1.9 for further evaluation (*Chang 2015*), retaining only autosomal and biallelic SNPs. Three different files were generated: .bed, .bim, and .fam. The .bed file is a binary file with the genotype calls. The .bim file contains extended variant information which has a row for each variant and information about chromosome, variant identifier, position in Morgans or centimorgans, base-pair coordinates, allele 1 and allele 2. The .fam file has sample information for each individual and information about family ID, within-family ID, within-family ID of father, within-family ID of mother, sex code and phenotype value. Variants with a missing rate higher than 2% were filtered out.

5.1.2.2 Kinship tests

To assess possible relatedness among the individuals in the dataset, the Kinship-based Inference for Genome-wide association study (KING) tool was

used (*Manichaikul 2010*), which can estimate kinship up to third-degree for arrays data. KING can be used through PLINK 2.0, which allows for the estimation of pairwise kinship coefficients.

Modern individuals from Donetsk oblast, Ukraine, were tested with KING and all were retained for further analyses as no relatedness was found (up to the third degree).

5.1.2.3 Principal component analysis (PCA)

PCA was performed with the software EIGENSOFT v.7.2.0 (*Patterson 2006*), using the *"smartpca"* program. Modern individuals which were coming from Donetsk oblast were analysed together with a published west Eurasian dataset (1469 individuals) of modern high-coverage sequences or genotypes acquired with Human Origin chip as in (*Capodiferro 2021*). For all modern individuals, analyses were done with pruned dataset and option *numotulieriter* set to 0 to disable outlier removal. The modern and ancient datasets were pruned in the plink format using PLINK v.1.9 (*Purcell 2007*) using the –indep-pairwise option. Parameters were set to 200 (indicates the window of SNPs in bp and calculates the linkage disequilibrium (LD) between each pair of SNPs in the window); 25 (shift for the new window); 0.4 (is for convergence criterion which is the threshold for stopping the algorithm based on log-likelihood improvement). In the end, the final dataset of modern published and studied in this research individuals reached 1514 and contained 409,084 SNPs.

PCAs were plotted with the ggplot2 package (*Wickham 2016*) using R version 4.3.0.

5.1.2.4 Admixture

A software program called ADMIXTURE (*Alexander 2009*) was used to estimate an individual's maximum likelihood of ancestry from multilocus SNP genotype datasets in which the individuals are unrelated. Admixture shows a gene flow between two ancestral previously separated individuals or populations in a new admixed individual/population. Through this process, patterns of genetic variation within and between individuals/populations are given an ancestry-based structure, which in turn affects the ability to infer demographic history, identify genetic targets of selection, and predict complex traits (*Gopalan 2022*).

The K values were set from 1 to 20, which represents the number of ancestral populations. For each K value ten repetitions were done. A log file with details about the run itself is also generated for every repeat; among other things, the CV (cross-validation) error value of every run was utilised to identify the K value with the lowest CV error value. One of the output files (Q files, which

contain an ancestry fraction) were used for clustering inference in the Pong tool (*Behr 2016*) v1.4.7. With the use of this tool, the best run out of the ten runs of for each K value was determined. This is achieved by folding the data (folds are distinct, non-overlapping subsets of the data), and using some folds for model training, and other folds for model validation. The model's performance is estimated by averaging the error over various folds. Usually, the value of K that has the lowest cross-validation error is selected. This suggests that the model will probably generalise well to fresh data containing that quantity of populations.

5.2 Ancient individuals

The archaeological specimens from seven ancient individuals analysed in this PhD thesis were obtained from our collaboration with the Archaeological Museum and the Department of Archaeology and Museum Studies, Faculty of History of Taras Shevchenko National University of Kyiv in face of Assoc. Prof. Pavlo Shydlovskyi and a Director of Archaeological Museum Liubov Samoilenko and a student Claudia Sharapova. We picked different fragments of bones or teeth of seven different individuals according to anthropological estimations. These individuals were found in four archaeological sites on the territory of modern Ukraine. They were originally gathered from archaeological excavations of burial mounds by a rescue expedition led by the Department of Archeology and Museum Studies between 1974 and 1994. The expedition, carried out by a renowned Ukrainian archaeologist, Prof. Dmytro Telegin (1919-2011), served as a base for student archaeological practice at the University Faculty of History. Remains belong to different cultures and periods of time: Yamna, Catacomb and Scythian.

For all but one of the analysed individuals listed in table 5.2.1, we were able to obtain the petrous part of the temporal bone, which is considered one of the best preservation sites for ancient DNA (aDNA) molecules (*Pinhasi 2015, Pinhasi 2019*). One specimen was represented only by two molars (UKR-b8). The bone fragments and teeth were processed in the ancient DNA facility of the Department of Biology and Biotechnology of the University of Pavia.

We studied three individuals from kurgan 21, which was located on a hill in the middle of a flat field surrounded by plantations, 1.8 km east of Pershotravneve village, in Nikopol raion of Dnipropetrovsk oblast (*Figure 5.2.1*, table 5.2.1). The kurgan, damaged annually, had a steep hemispherical shape with a flattened top. It was constructed in the Yamna period in three phases, containing ten burials: eight from the Yamna culture (one of which UKR-b8, *Figure 5.2.2.A*), one from the catacomb (UKR-k21b-10A, UKR-k21b-10b, *Figure 5.2.2.B*), and one Scythian.

The burial #8 had (UKR-b8, *Figure 5.2.1.A*) a grave filled with chernozem mixed with mainland clay. The vertical-walled pit measured 2.36x1.74 m and was 4.72 m deep. Plant decay, likely from a woven mat, and fragments of wood were found in the grave. The adult buried lay on the left side, head facing northeast, with poor bone preservation. Red ochre traces were noted on the skull and throughout the grave, but no artefacts were found.

Information on the burial #10 (with UKR-k21b-10A, UKR-k21b-10B, *Figure 5.2.2.B*) says that it was located 22 m southwest of the kurgan's centre. This burial had two adult skeletons (A and B) lying flat on their backs, heads to the north. Bones were poorly preserved and displaced, likely due to a vault

collapse. Skeleton A had disorderly bones, with the lower jaw resting on the femur and the skull damaged. Skeleton B's skull was densely filled with chernozem and dark red ochre. Articulated animal bones were found near the eastern wall. This burial had no accompanying artefacts (*Бондарь 1978*).

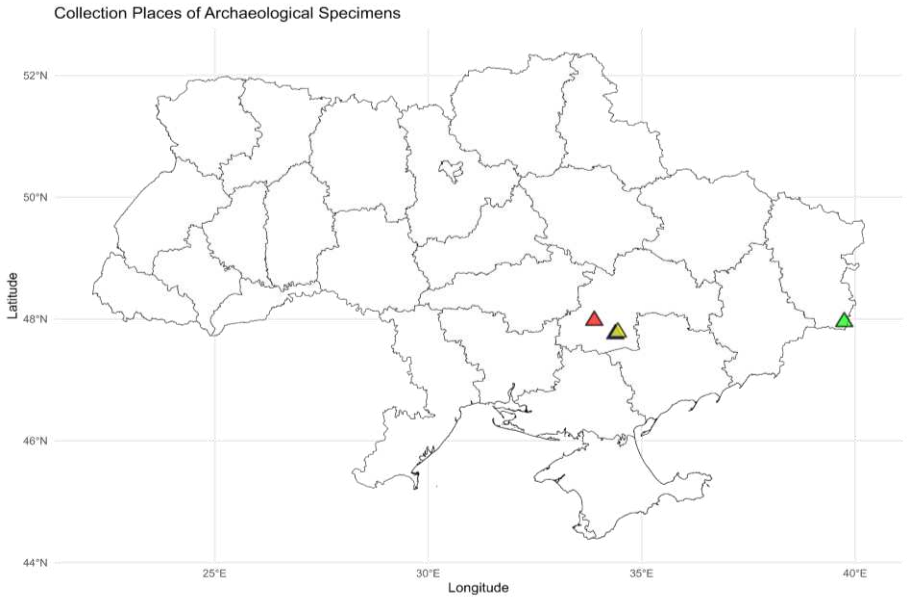


Figure 5.2.1. *Map of Ukraine with four collection places represented by dots of different colours according to each site code (see Table 5.2.1)*

Table 5.2.1. *List of ancient individuals*

Sample ID	Collection place	Site code	Culture	Skeletal element	14C (BP)
UKR-K4b13	Krynychne, Dovzhansk raion, Luhansk oblast	01	Yamna	Petrous	4165±21
UKR-b8	Pershotravneve, Nikopol raion, Dnipropetrovsk oblast	02		Molars	4194±21
UKR-k21b10A	Pershotravneve, Nikopol raion, Dnipropetrovsk oblast	02	Catacomb	Petrous	*NA*
UKR-k21b10B				Petrous	3891±21
UKR-K4b4	Novoivanivka, Nikopol raion, Dnipropetrovsk oblast	03	Catacomb	Petrous	3928±21
Ukr-K1b1.1	Novovitebske, Kryvyi Rih raion, Dnipropetrovsk oblast	04	Scythian	Petrous	2203±19
Ukr-K1b1.2				Petrous	2179±19

NA - data not available

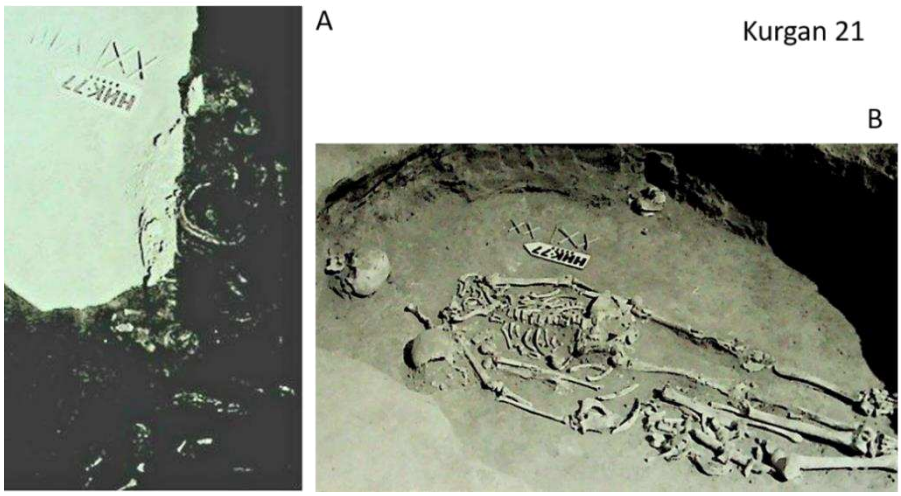


Figure 5.2.2. *Individuals UKR-b8 (A), and UKR-k21b-10A, UKR-k21b-10B (B) (Бондарь 1978)*

Kurgan 4 of burial with individual UKR-K4b13 (*Figure 5.2.3*) was located 3.35 km south of the village Krynychne in Dovzhansk raion, Luhansk oblast. It was discovered at a construction site for the Dovzhan irrigation system, it is the easternmost point of Ukraine, close to the Russian border (*Figure 5.2.1*, table 5.2.1). The area is currently under Russian occupation.

Figure 5.2.3. *Individuals UKR-K4b13, kurgan 4 (Бондарь 1978)*

The kurgan had a hemispherical shape with a diameter of up to 25 m and was preserved to a height of 1.7 m. In ancient times, it was covered with small rough stones, likely protecting it from plowing. The ancient horizon was recorded 1.5 m below the surface, and the chernozem beneath the mound

was 0.6-0.65 m thick. Here were found individual UKR-K4b13, belonging to Yamna culture. The deceased was an adult, lying on his back with his legs bent at the knees, which fell to the right, his head to the West. The skull was turned on the left temple. The right hand is stretched along the body, the left hand is bent at the elbow and placed on the stomach. Bone preservation was satisfactory. The bone of the buried person and the bottom of the burial were abundantly sprinkled with ochre. No inventory was found in the grave.

The representative of Catacomb culture (UKR-K4b4) was found in Novoivanivka, Nikopol raion, Dnipropetrovsk oblast, which is very close to Pershotravneve (*Figure 5.2.1*). The kurgan also has number 4. The kurgan was plowed annually and damaged by burrows, but was preserved to a height of 0.63 m with a diameter of 30 m. Buried chernozem was found at a depth of 0.58 m, reaching a thickness of 0.35 m. Six burials were uncovered: three from the Yamna culture, two from the Catacomb culture (one of which individual UKR-K4b4), and one from the Zrubna culture. Burial 4, located in the southwestern sector of the kurgan, was in a catacomb and measured 2.14x1.5x1.38 m, though heavily damaged by fox holes. The skeleton was largely destroyed, except for a few bone fragments, finger phalanges, and a part of the skull found in the northeastern corner of the chamber. No grave goods were discovered.

Remains of Scythian individuals from the Early Iron Age (UKR-K1b1.1 and UKR-K1b1.2) were found in kurgan 1 (primary of a catacomb burial), which was located 4.25 km southwest of Novovitebske village of Kryvyi Rih raion, Dnipropetrovsk oblast (*Figure 5.2.1*, Table 5.2.1), the kurgan had been damaged by agricultural activity, surviving to a height of 1.13 m with a 34 m diameter. Amphora fragments, possibly from a winepress, were found on the surface. The entrance, partially collapsed and looted, revealed scattered human bones from two adults, along with bull bones, pottery fragments, and pieces of an iron scale armour. Based on a mix of skeleton fragments and bones, it was assumed, there were two Scythian individuals a male and a female (*Бондарь 1977*).

Geographical location of the burials (*Figure 5.2.1*) makes these individuals unique and precious, since eastern and central-southern Ukraine is massively destroyed in the Russian-Ukrainian war, which began in its full scale during my PhD on February 24, 2022. Some of these areas are still occupied and there is much evidence of destruction and robbery of archaeological and historical museums of Ukraine by Russian forces.

5.2.1 Radiocarbon dating

Radiocarbon dating, or ^{14}C dating, is a widely used dating technique in environmental and geosciences studies as well as archaeology. The technique can be used to determine the age of a range of organic materials back up to 50,000 years. We performed ^{14}C dating on outsourcing in Germany, Mannheim in Curt-Engelhorn-Zentrum Archäometrie gGmbH facilities (*https://ceza.de/english/cezapedia/methods/c-14-dating*).

Individuals, selected for the radiocarbon analysis based on endogenous DNA content following preliminary sequencing results, were the following: from Yamna culture UKR-K4b13 and UKR-b8; from Catacomb culture only UKR-k21b10B; from Scythians both Ukr-K1b1.1 and Ukr-K1b1.2.

According to the report received on 12.07.2024 (#240241), for our sample preparation collagen was extracted from the bone or teeth by modified Longin method, purified by ultrafiltration (fraction > 30kD) and freeze-dried. The remaining sample material is combusted to CO_2 in an Elemental Analyzer (EA). CO_2 is then converted catalytically to graphite. ^{14}C analyses were realised on a MICADAS-type Accelerator Mass Spectrometry system with following normalisation and calibration.

5.2.2 Ancient DNA laboratories

All work with ancient individuals, up to the preparation of the libraries for sequencing, was carried out in the ancient DNA Facility at the Department of

Biology and Biotechnology of the University of Pavia (*Figure 5.2.2.1*). Bench work was performed in a dedicated clean area: the Boone Room and the Molecular Room. To reduce the risk of contamination in this area, a procedure of appropriate laboratory clothing is used. To access the clean area, operators must wear a mask, bouffant cap, shoe covers, and gloves while still in the first room. Once in the cloak room, they must put on a gown, shoe covers, and an extra pair of long gloves. The procedure must be completed by cleaning the outer clothing with DNA-Exitus Plus™ or other detergents designed to remove DNA contamination from surfaces.

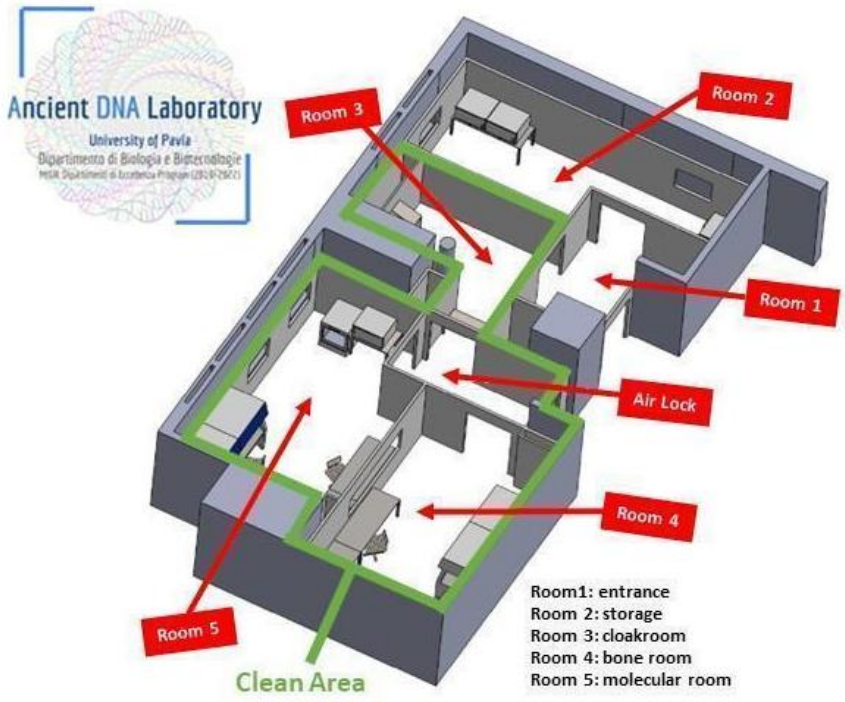


Figure 5.2.2.1. *Schematic representation of the aDNA facility at the University of Pavia*

5.2.3 Ancient DNA extraction

A silica-based method was used for aDNA extraction that allows recovery of very short DNA fragments (*Pinhasi 2019, Damgaard 2019, Harney 2021*). The DNA extraction process takes two days and involves multiple steps.

5.2.3.1 Sample decontamination

Given the prevalence of soil contamination and the use of museum glue in archaeological specimens, a decontamination process is essential prior to DNA extraction. Typically, bone fragments or teeth are received at the laboratory in plastic bags, which must be cleaned with DNA-Exitus Plus and alcohol before being introduced into the clean rooms.

Once the samples are brought to the bone room, the bags containing them must be placed in the short-wave (254 nm) UV crosslinker for five minutes on each side. Archaeological specimens may be removed from the bags and placed in a prepared sterile plastic cuvette. Large visible soil particles can be

removed manually at this stage using sterile forceps or spatulas. Later, a wash in distilled water can be performed to remove any remaining dirt and broken fragments. If the examined fragments still appear soiled, they may be washed in 1:5 bleach, then rinsed rapidly in H_2O, then in isopropanol to accelerate drying and remove any remaining water; each step may take up to 30 seconds. If radiocarbon dating of the bone is expected, this last step should be skipped, or a portion of the bone may be cut for this purpose before the isopropanol wash. The specimens are dried in a cross-linker for 3 minutes on each side, then cleaned with sandpaper if necessary. Finally, the specimens are dried in the UV cross-linker for up to 5 minutes on each side. After these steps, the bones are ready for drilling or further digestion.

5.2.3.2 Temporal bone drilling and powdering

After the decontamination process the petrous bone can be processed in two main ways: 1) direct drilling for obtaining sample powder or 2) pre-digestion followed by powdering of the fragment.

The hood used for cutting and drilling must be cleaned with DNA-Exitus Plus and alcohol and covered with aluminum foil at the bottom. Cleaning and changing the foil is critical before each new sample is placed under the hood to prevent cross-contamination. To access the inner portion of the petrous bone where the cochlea is located, which is considered one of the best sources of aDNA preservation bones due to its protected inner location in the scull, a diamond disk attached to a Dremel tool is used to cut through the surrounding spongy bone. The cochlea is cut perpendicularly and half of it is used to make bone powder. The prepared fragments are placed then in zirconium oxide grinding jars of the Retsch MM400 mixer mill. After grinding, the powder is collected with a spatula in sterile tubes, and 50-100 mg of the powder is taken for the following digestion process, while the rest is stored at -20 °C.

In case of predigestion, first, selected fragments are placed in falcons or 2 mL tubes in a predigestion solution (1mL EDTA + 50 µL of proteinase K (20 ng/µL) + 50 µL of 10% N-lauroylsarcosine) for 1 or 1.5 hours at 37° C. The bone fragments obtained from the pre-digestion must be dried well for at least 30 minutes on each side, only then they can be powdered as described in the previous paragraph.

5.2.3.3 Teeth drilling and powdering

Teeth are one of the best parts of the human body for preserving ancient DNA (*Dabney 2013*). Several techniques can be used to extract DNA from such remains, including whole tooth root (WTR), Damgaard-style cementum sampling (D-style), and minimally destructive extraction (MDE).

The extraction method we routinely use in our laboratory is WTR (*Posth 2016*), which requires cutting and powdering the entire root of each tooth. After the surface of the teeth was cleaned, we proceeded to cut the crown from the teeth using a Dremel drill with a diamond wheel. Finally, the root was powdered using the Retsch MM400 machine. The powder was collected and part of it was used for the next step of digestion, the rest was kept in the freezer at -20°C.

5.2.3.4 Minimally destructive extraction from teeth

Minimally Destructive Extraction (MDE) was performed on the UKR-b8 specimen on both molars we received. This method allows the sample to be preserved without drilling or grinding. This makes it useful for other analyses such as radiocarbon dating or morphological studies (*Harney 2021*). The biological target of this protocol is the thin layer of cementum on the outside of the tooth root, which is composed of multipotent stem cells. The cells within the cementum matrix, cementocytes, can secrete cementum to stabilise the teeth in their sockets (*Lindsay 2013*). Cementum shares some similarities with petrous bone in that cementocytes can retain DNA like osteocytes. However, cementum does not undergo structural changes in the same way as the cochlea, but accumulates throughout life (*Pinhasi 2015, Harney 2021*). According to estimates of DNA preservation in ancient specimens, dentin has a lower content of endogenous DNA than cementum (*Damgaard 2015*). According to the MDE protocol we proceeded with the following steps:

- Physically clean the tooth: the tooth surface was cleaned with a 2% bleach solution until visible stains were removed. We used isopropanol to remove the bleach.

- UV irradiation: the tooth was dried with UV in a cross-linker for five minutes at 254 nm on both sides. It is important that the sample is dried.

- Crown protection: we covered the crown and root tip with parafilm to protect these parts from the digestion reagents used in the next steps (*Figure 5.2.3.4.1*). In this step, it is important not to apply too much pressure, as the old tooth may crack.

- Digestion: The tooth was placed in a 5 mL tube with the exposed part of the root facing down. The tube was filled with 1 mL of extraction buffer and digested for 2.5 hours at 37°C with gentle vertical rotation to allow the buffer to circulate around the root.

Figure 5.2.3.4.1. *Example of a tooth that has been wrapped in parafilm, leaving only the lower part of the root exposed. Note the tail of parafilm present to allow for easier handling of the tooth*

5.2.3.5 Sample digestion

Approximately 50-100 mg of powder (petrous bones or teeth) was digested with a lysis buffer consisting of EDTA solution pH 8.0 (0.5 M) with 50 µL of 20 mg/mL Proteinase K and 50 µL of 10% N-laurylsarcosine for 24 hours with constant inclined rotation at 37°C in a 2 mL Eppendorf LoBind tube. The reagents and their concentrations are listed in the table 5.2.3.5.1.

Bivalent metal ions are sequestered by the chelating effect of EDTA. By inhibiting metal-dependent enzymes such as nucleases, this ability stops DNA degradation. Cell proteins are degraded, and peptide bonds are hydrolysed by Proteinase K. The addition of a reducing chemical (N - Lauroylsarcosine) causes cell membranes to rupture and accelerates protein degradation. A negative control was included in this step to check for reagent contamination. Parafilm was used to seal the tubes, and the powder was vortexed to suspend it. Continuous oblique rotation at 15 revolutions per minute (rpm) was performed during an overnight incubation at 50°C.

Table 5.2.3.5.1. *Reagents used for digestion during the first day of DNA extraction*

Reagent	**Description**
Lysis buffer	1 mL of 0.5 M EDTA pH 8.0 50 µL of 20 mg/mL proteinase K 50 µL of 10% (g/mL) N-lauroylsarcosine

5.2.3.6 Ancient DNA purification

The second day of extraction takes place in the molecular room. It involves purification of the extracted DNA using the Qiagen MinElute PCR Purification Kit.

The kit contains:

- PB Buffer, which contains a high concentration of guanidine hydrochloride and isopropanol and is used to denature proteins and promote DNA binding to the silica membrane in the spin column, allowing DNA to adhere to the membrane while impurities do not.

- PE Buffer, which is diluted 5X in absolute ethanol; it helps to wash away residual impurities that may still be bound to the silica membrane. Ethanol helps to keep the nucleic acids bound to the silica membrane while washing away impurities.

- EBT Buffer, which consists of 10 mL EBT Buffer (10 mM Tris-HCl, pH 8.5) and 5 µL Tween 20 (final concentration: 0.05%). Buffer EBT is a low salt buffer designed to facilitate the release of nucleic acids from the silica membrane.

Samples that have undergone digestion should be centrifuged to precipitate undigested particles. The supernatant should be mixed with 9 mL PB in a 15 mL Falcon tube and incubated for 10-15 minutes at room temperature. During incubation, the Zymo Research Reservoir can be attached to the MinElute silica membrane columns and placed in a 50 mL Falcon tube. The solution consisting of PB and supernatant is then applied to the MinElute column. During centrifugation, nucleic acids bind to the silica membrane of the columns. Therefore, the supernatant is discarded.

Then the DNA has to be washed from contaminants and proteins by two washing steps with the alcohol-based PE buffer. For this purpose, the MinElute columns are transferred into 2 mL sterile tubes. After washing, the columns are dry-spun by centrifugation at 16,400 g. Dried columns are transferred into 1.5 mL Eppendorf LoBind tubes.

Finally, purified DNA is eluted in 50 °C preheated EBT buffer. First, 52 µL of EBT is added to the centre of the column, incubated and centrifuged (at 3,500g for 1 min). The second step is repeated in the same way but with 23 µL EBT. After spin-drying at 8,000g for 1 min, the extracted DNA is safely stored at -20 °C.

5.2.3.7 Ancient DNA quantification and quality checks

Quantification of the DNA extracts (2 µL) is performed in the modern laboratory on the Qubit 4™ Fluorometer. Accurate quantification from 5 pg/µL to 120 ng/µL is possible with the Qubit 1X dsDNA High Sensitivity (HS) Assay Kit, which has a detection range of 0.1-120 ng dsDNA in 1-20 µL of sample. The fluorescent dye (PicoGreen dye) contained in the kit binds selectively to double-stranded DNA (dsDNA) and its fluorescence intensity is directly proportional to the amount of dsDNA present in the sample. Calibration of the Qubit instrument is performed using Standard 1 and Standard 2 (10 µL each) with the 190 µL of ready-to-use working solution. After calibration, quantification of the samples is performed in this step by adding 1 µL of DNA to 199 µL of working solution in new clear 0.5 mL PCR polypropylene tubes. After centrifugation and vortexing for five seconds, the tubes are left at room temperature for approximately five minutes. It should be noted that the negative control must be quantified to screen for potential contamination.

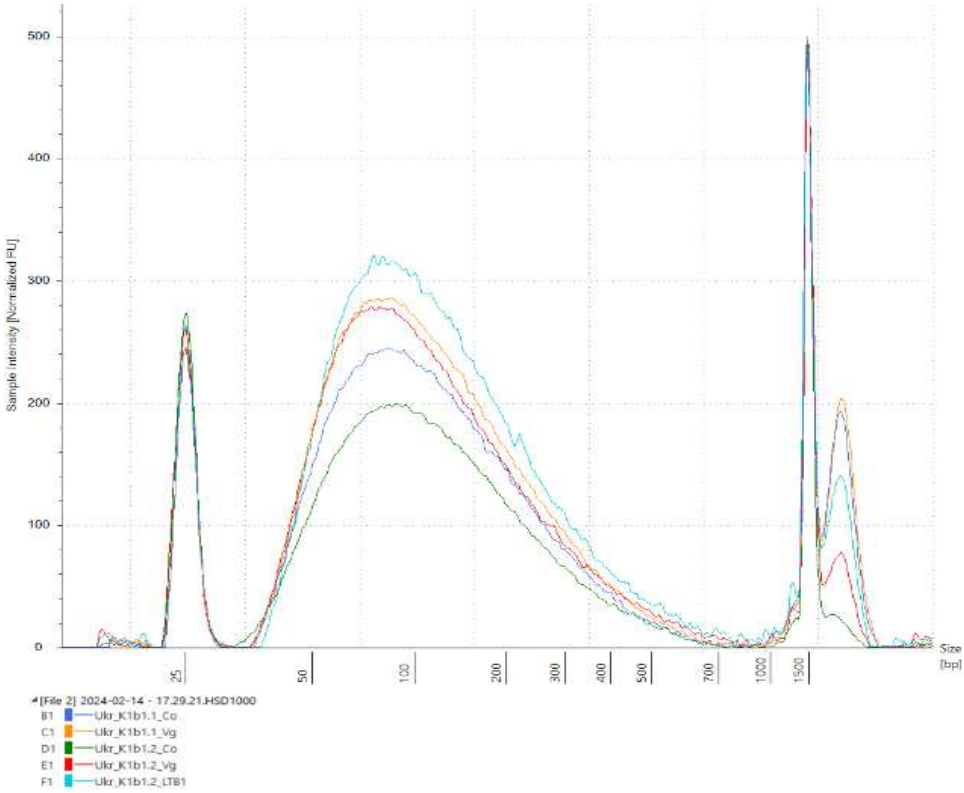


Figure 5.2.3.7.1. *Tape station analysis of different extracts from the petrous bones of individuals UKR-K1b1.1 and UKR-K1b1.2. Co stands for Cochlea, Vg stands for Vestigial channels, LTB means Lateral Temporal Bone*

All extracted DNAs were screened by capillary electrophoresis using the Tape Station 4150 (Agilent, USA) with High Sensitivity Kit. Capillary electrophoresis checks the quality and fragment size of the extracted DNA. The average size of the fragments should be less than 200 base pairs (*Figure 5.2.3.7.1*).

5.2.4 Ancient Library preparation

All extracted DNAs with a concentration higher than the blank and a suitable pattern from the Tape Station analysis undergo library preparation. This process takes two days and is performed entirely in the molecular room.

For library preparation, the protocol of half-UDG treatment was used (*Meyer 2010*, *Rohland 2015*). The protocol involves the use of USER enzymes designed for the treatment of C to T (or G to A) nucleotide substitutions. Such damage is a result of hydrolytic deamination that occurs in ancient DNA. Cytosine is a pyrimidine that loses its amino group more quickly and easily than other types of nucleobases. After deamination, it becomes uracil, which by the complementary rule is paired with adenine during the replication cycles. In another chain, after replication, it brings thymine to the actual place of cytosine. Uracil-DNA glycosylase is able to reduce the amount of this substitution by catalysing the hydrolysis of the N-glycosidic bond to release Uracil. For downstream analysis, UDG treatment significantly reduces the rate of ancient DNA errors. However, libraries without UDG treatment have an advantage in damage pattern analysis for endogenous read detection. The ancient DNA error rate is reduced by half of the UDG treatment, but the damage signal at the terminal bases of the molecules is still detectable and allows the validation of ancient DNA reads (*Rohland 2015*).

5.2.4.1 Partial UDG treatment

A mixture containing USER enzymes should be prepared using the components listed in Table 5.2.4.1.1.

Table 5.2.4.1.1. *Reagents used for partial UDG*

Reagents	**Volume (µL) per sample**
Tango Buffer (10X)	5
dNTPs (10mM each)	0.5
ATP (100 mM)	0.5
USER (1U/µL)	3.6

The amount of DNA to be added should be between five and 15 ng, as follows: 35 µL if concentration (C) of the DNA <0.2; 23 µL if C <0.5; 15 µL if C<1; 5-10 µL if C>1. The final amount of DNA must be 35 µL, the missing volume is made up with water. It is crucial to follow this balance of DNA input, because too low quantity of aDNA can result in too high Uracil removal, this fragmentation of aDNA and making it not accessible for libraries preparation, in case of too high input the reaction enzymes can be overloaded and do not work sufficient, leaving Uracil on place. A Nuclease Free Water (NFW) negative control is also included.

Incubation for 30 minutes at 37°C is needed to activate the enzyme. Then an incubation for 2 min at 12°C stops the hydrolysis of N-glycosidic bonds.

5.2.4.2 UDG inhibition

To prevent any further work of USER enzymes 3.6 µL of uracil glycosylase inhibitor (UGI) is added to each tube. The 9.5 kilodalton (kDa) UGI protein from *Bacillus subtilis* bacteriophage PBS1 dissociates UDG-DNA complexes and inhibits UDG binding with a 1:1 stoichiometry. The samples are quickly centrifuged and incubated in a heat cycler at 37 °C for 30 minutes after being mixed by flicking the tube with a finger. This is followed by 2 min at 12 °C.

5.2.4.3 Blunt-End Repair

This step requires the use of T4 Polynucleotide Kinase (which is capable of phosphorylating the 5' of DNA for ligation, labelling the ends, and removing 3' phosphoryl groups) and T4 DNA Polymerase (which catalyses the synthesis of DNA in the 5'→ 3' direction and fills in the gaps, removes 3' overhangs, or fills in 5' overhangs to form blunt ends). For preparation of the highly fragmented DNA these reagents are used in the proportions described in the Table 5.2.4.3.1.

Table 5.2.4.3.1. *Reagents used for blunt-end repair*

Reagents	**Volume (µL) per sample**
T4 PNK (10 U//µL)	2.5
T4 polymerase (5 U/µL)	1

The samples are mixed by flicking the tube with a finger, rapidly centrifuged and incubated in a thermal cycler at 25 °C for 20 min, then in another thermal cycler at 12 °C for 10 min.

5.2.4.4 Sample clean-up

After this treatment, we proceeded to the first purification step. This was performed using the QIAQuick Purification Kit according to the standard protocol. Each end-repaired DNA sample is mixed with 500 µL PB Binding Buffer and incubated for 2 min at room temperature in a QIAQuick column.

500 µL of PB Binding Buffer are added to each end-repaired DNA sample, which is mixed and then incubated for 2 min at room temperature on a QIAQuick column. The columns are then centrifuged in an Eppendorf MiniSpin for 1.5 min at 7000 rpm. As the DNA binds to the silica-based membrane, the flow-through is discarded. The columns are then washed with 600 µL of PE buffer, centrifuged at 3400 g for 1 min and then centrifuged again at maximum speed to dry-spin. The columns are transferred to a new 1.5 mL Eppendorf LoBind tube and the DNA is eluted from the silica membrane by adding 32 µL of pre-heated (50 °C) EBT buffer and incubating for 10 min at room temperature. Samples were centrifuged after this step for 1 min at 7000 rpm on Minispin plus 1 min at 11000 rpm.

5.2.4.5 Adapter ligation and second clean-up

The adapter ligation step is done with T4 DNA ligase which joins blunt end and cohesive ends and fix single stranded nicks in duplex DNA. Adapters are needed for creation of binding sites for primers of subsequent amplification and for recognition of DNA fragments by sequencing platform. A master mix is prepared according to proportions from table 5.2.4.5.1.

Table 5.2.4.5.1. *Reagents used for adapter ligation.*

Reagents	**Volume (µL) per sample**
T4 DNA ligase buffer (10X)	4
PEG-4000 (50%)	4
Adapter mix (10 µM each)	1
T4 DNA ligase	1

Ten µL of the mix is applied to each sample, mixed by pipetting, and incubated for 30 minutes at 22°C. At this point, a further clean-up is required to remove adapters in the residue. This is done using the QIAQuick Purification Kit as described in the previous section. This time the amount of PB buffer is increased to 600 µL, while the rest of the protocol remains the same. After this process, the samples could have a safe freezing point for up to two weeks.

5.2.4.6 Adapter fill-in

Adapter fill-in is performed with Bst polymerase, which has 5'→ 3' polymerase activity but lacks 5'→ 3' exonuclease activity (*Kircher 2012*).

A reagent mixture is prepared according to Table 5.2.4.6.1. Samples are mixed by pipetting 10 µL of the mixture and incubated at 37°C for 30 minutes followed by incubating at 80°C for 10 minutes.

Table 5.2.4.6.1. *Reagents used for adapter fill-in*

Reagents	**Volume (µL) per sample**
Thermopol reaction buffer (10X)	4
dNTPs (2.5 mM each)	4
Bst polymerase, large fragment (8 U/µL)	2

To move forward to the amplification step, the libraries have to be quantified with Qubit 4 Fluorometer and Qubit 1X dsDNA HS Assay kit, as described in paragraph 5.2.4.7.

5.2.4.7 Library enrichment and indexing

Finally, the libraries should be amplified with the Platinum SuperFi™ II PCR Master Mix using MPI and LP indexed primers as described in (*Kircher 2012*). Indexes must be selected with a unique combination of barcodes for sample identification after sequencing. The amount of adapter-ligated DNA fragments added can vary from 2 to 8 µL, resulting in a final concentration in the PCR tube of 1-1.5 ng. NFW is added to make up the remaining volume. Reaction components are added as described in Table 5.2.4.7.1. Final volume of the PCR reaction is 20 µL.

Table 5.2.4.7.1. *Reagents used for amplification*

Reagents	**Volumen (µL) per sample**
2X Platinum™ SuperFi™ II PCR	10
P7 index primer MPI (10 µM)	0.5
P5 index primer LP (10 µM)	0.5

PCR mix with added DNA is taken out of the sterile area and amplified in a dedicated thermocycle using the program in table 5.2.4.7.2.

Table 5.2.4.7.2. *PCR program for amplification*

Step	Temperature (°C)	Time	Cycles
Initial denaturation	98	30 seconds	1
Denaturation	98	10 seconds	12
Annealing	67	10 seconds	
Extension	72	5 seconds	
Final extension	72	5 minutes	1
Hold	4	∞	

5.2.4.8 Library clean-up

After the amplification, the libraries were quantified with the Invitrogen Qubit fluorometer. The best libraries were eventually purified with the QIAQuick PCR purification kit as described in paragraph 5.2.4.4, or with Illumina DNA Prep Purification Beads (IPB). Bead cleanup is required if short dimers or long fragments (which can derive from modern DNA contamination) are detected. The type of purification should be selected based on the contamination present and can be done in three different ways: Left Size Selection (1.2x);

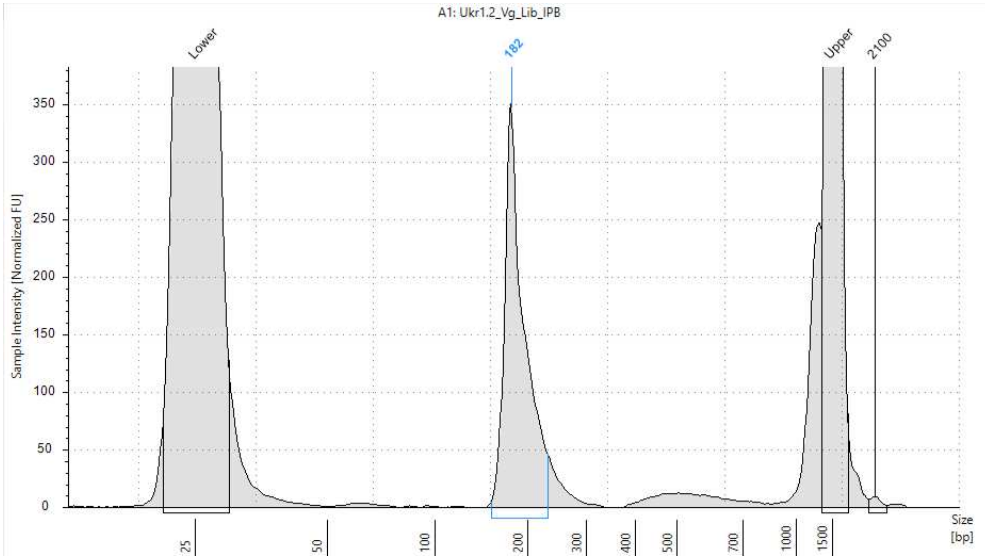


Figure 5.2.4.8.1. *Example of electropherogram (TapeStation) for library UKR-K1b1.2 processed with IPB purification. Fragment size is indicated on the horizontal axis and the intensity of the fluorescent signal on the vertical axis. The peak between 400-700 might be represented by modern DNA contamination, therefore on this step we also select the best samples for further sequencing. The peak of 182 bp is compatible with aDNA fragment size*

Right Size Selection (0.7-1.5x); or Double Size Selection (0.7-1.2x). In all cases, the DNA volume was brought to 80 µL with NFW.

Purified libraries were eluted with 35 µL of preheated EBT buffer pre-warmed at 50°C (QIAQuick) or 32 µL RSB from Illumina (Illumina DNA Prep Purification Beads). Quality control of the purified product was performed on the Agilent TapeStation using the D1000 kit, as it covers the range of 35 bp to 1000 bp with a concentration between 0.1 and 50 ng/µL (*Figure 5.2.4.8.1*).

5.2.5 Preparation for shotgun sequencing on Illumina platform

Prepared libraries were sequenced in different runs on Illumina platforms using paired-end and for some trials single-end sequencing after normalisation and pooling steps.

5.2.5.1 Library normalisation and pooling

Different libraries should be normalised to achieve uniform coverage across different samples and/or fragments. If this step is skipped, overrepresented libraries may consume a disproportionate amount of sequencing reads.

The libraries were normalised to either 2nM or 4nM. Normalisation considers the average size of the DNA libraries estimated on the TapeStation and the concentration measured on the Qubit 4 Fluorometer system to calculate the molarity of each dsDNA library using the following formula:

$$Concentration\ in\ nM = \frac{Concentration\ in\ ng/ul}{660\frac{g}{mol} \times Average\ library\ size\ in\ 150-750\ bp\ region} \times 10^6$$

660 g/mol represents the average molar mass of a single base pair (*Duewer 2018*)

Each library was diluted to the final concentration with Illumina Resuspension Buffer (RSB) according to the following formula:

$$Vlibrary = \frac{Vfinal \times Final\ pool\ concentration\ nM}{Library\ concentration\ nM}$$

$$V_{NFW} = V_{final} - V_{library}$$

Volume of each library in the final pool was calculated using the formula:

$$Vlibrary\ to\ be\ pooled = 3 \times \left(\frac{Final\ pool\ concentration\ in\ nM}{Normalised\ library\ concentration\ in\ nM}\right)$$

The final molarity of the pool was checked by quantification on Qubit 1X dsDNA HS Assay kit and through the TapeStation analysis.

5.2.5.2 Pool denaturation and dilution

The pool is diluted and denatured in accordance with "Protocol A: standard normalisation method" as outlined in the Illumina handbook for the NextSeq 550 and NextSeq 500 sequencing systems. Table 5.2.5.2.1 lists the reagents needed for pool dilution and denaturation.

Table 5.2.5.2.1. *Reagents used for pool denaturation and dilution*

Reagents	**Description**
0.2 N NaOH	98 µL NFW + 2 µL 1 N NaOH
200 mM Tris-HCl (pH 7.0)	
HT1 hybridization buffer (Illumina)	High-salt concentration buffer used to dilute denatured libraries
PhiX control v3 (Illumina)	Diluted with HT1 to 1.5 pM or 1.8 pM) according to the Illumina kit used

For pool denaturation, 0.2 N NaOH solution is added to the pool in the following volumes (Table 5.2.5.2.2) and incubated for 5 minutes at room temperature.

Table 5.2.5.2.2. *Volume of 0.2 N NaOH to be added to the denature pool*

Starting library concentration	**Library (µL)**	**0.2 N NaOH (µL)**
4 nM	5	5
2 nM	10	10

To dilute the denatured pool 200 nM Tris and prechilled HT1 are added in turn to obtain a final concentration of 20 pM (Tables 5.2.5.2.3 and 5.2.5.2.4).

Table 5.2.5.2.3. *Volumes of Tris-HCl to hydrolyse the NaOH*

Starting library concentration	**200 mM Tris-HCl pH7 (µL)**
4 nM	5
2 nM	10

Table 5.2.5.2.4. *Volumes of HT1 to dilute the pool*

Starting library concentration	**Prechilled HT1 (µL)**
4 nM	985
2 nM	970

The flow cell to be utilised for sequencing determines the library pool's final dilution. To get the loading concentration of 1.8 pM for high output kits, mix 1183 µL of prechilled HT1 with 117 µL of denatured library solution. The loading concentration for mid output kits should be 1.5 pM, which can be accomplished by diluting 1203 µL of prechilled HT1 with 97 µL of denatured library solution. In both situations, the total volume is 1.3 mL. Lastly, as a sequencing control, an Illumina Phix spike-in equal to about 1% of the final pool is introduced.

5.2.5.3 Illumina Sequencing

Most of the ancient libraries were sequenced on the Illumina NextSeq 550 platform on mid- or high-output flow cells with a run of 75 to 150 cycles, since the ancient DNA is highly degraded and its size is often under 100 bp. A few last runs were performed with the Illumina NextSeq 2000 on flow cell cartridges P2 and P3, with a run of 200 cycles, pair end sequencing (2x100).

Illumina NGS systems use a sequencing by synthesis (SBS) method (*Figure 5.2.5.3.1*). It starts with the clustering process, where the prepared DNA anneals to the oligos ("P5" and "P7") located on the flow cell, which are complementary to one of the adapters of the library chains. The first amplification fixes the DNA fragments attached to the flow cell and allows the bridge amplification to begin after the denatured DNA strands are separated and the original template is washed away. The amplification process is carried out by DNA polymerase and takes place simultaneously in millions of clusters, resulting in massive clonal amplification of the DNA fragments. After the clusters are filled with copies of libraries, the reverse strands are removed from the clusters. Sequencing begins with primer fixation and extension of the first nucleotides. Only one of the four fluorescence-labelled dNTPs is incorporated into the new strand at a time. Each fluorescent labelled nucleotide also has a terminator sequence that allows the reaction to be stopped. This gives the Illumina sequencer's optical system time to read the light signal and define the nucleotide added to the growing strand. After the terminator sequence is removed, the polymerase can continue its work at the next nucleotide position (*https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html*).

Illumina sequencing has high accuracy, especially in homopolymers, a high yield of error-free reads, and a high percentage of base calls above Q30, which means that the probability of an incorrect base call is 1 in 1000, or 0.1% (*Bentley 2008, Ross 2013*).

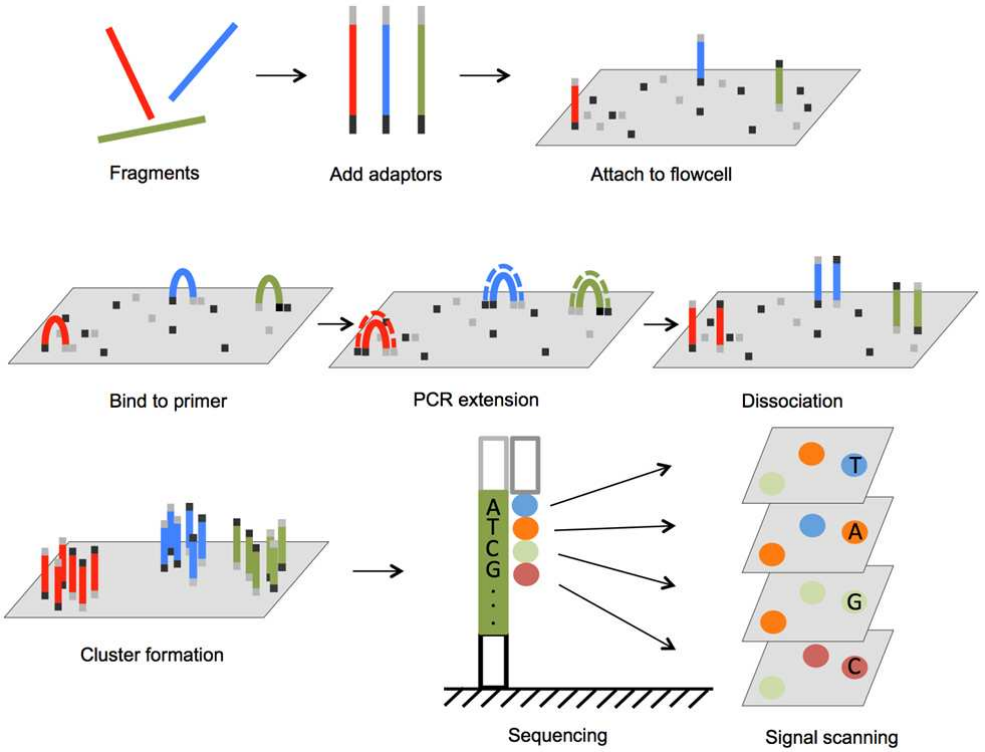


Figure 5.2.5.3.1*. Illumina sequencing process overview* (*Lu 2016*)

The ancient libraries were sequenced on the Illumina NextSeq® 550 System at the IRCCS Mondino Foundation of Pavia (Drs. Stella Gagliardi and Rosalinda Di Gerlando) and on Illumina NextSeq® 2000 System at the Dept. of Molecular Medicine (Dr. Elisa Giorgio).

Medium output cartridges were used for pre-screening of endogenous yield. If the endogenous yield was greater than 5%, the libraries were re-sequenced on high output cartridges to achieve a final depth of coverage of 0.5x.

5.2.6 Ancient data analysis

Ancient DNA has to be analysed in a modified way compared to modern DNA. A workflow includes processing of raw data retrieved after sequencing, alignment to a reference genome and estimation of the endogenous content of ancient DNA in a sample. Important step is evaluation of the possible modern human DNA contamination (*Orlando 2021*).

5.2.6.1 Demultiplexing

Demultiplexing of the libraries sequenced on the Illumina NextSeq 550 system and NextSeq 2000 has the same workflow. It is a process of identification and assortment of the raw data from Illumina systems in Binary Base Call (BCL) format.

Libraries which were sequenced with Illumina NextSeq 550 system and Illumina NextSeq 2000 were demultiplexed from BCL format, based on the detection of dual indexes that were added to DNA fragments during library preparation. The conversion from binary files into FASTQ files is done with Illumina software bcl-convert v.4.2.4 (*https://support.illumina.com/sequencing/sequencing_software/bcl-convert/downloads.html*).

All parameters were used in default, except for "--no-lane-splitting true", which is used to put together in one file the FASTQs from different lanes. FASTQ files are text files that contain the information about sequence identifier, the nucleotide sequence, separator and a base call quality scores, data from the filtered clusters on a flow cell. As a result of this demultiplexing the software generates one or two FASTQ files: Read 1 (R1) and Read 2 (R2, if it was a paired-end run).

5.2.6.2 Quality check

A Unix-based pipeline created by our research group (with assistance from Drs. N. Rambaldi Migliore and M. Capodiferro) was used to evaluate the two raw FASTQ files. The program fastQC v0.11.8 (*Andrews 2012*) is used to check the quality of the raw reads. In order to determine whether there are any problems with the data before performing additional studies, it provides a modular collection of analyses.

5.2.6.3 Processing of raw reads

The software AdapterRemoval v2 was used to remove the adapters from the DNA sequences and for trimming low-quality bases. In brief, the minimum base quality score was set as 2 from the 5'/3' ends of reads (*Lindgreen 2012, Schubert 2016*). Ambiguous bases (N) from 5'/3' ends of reads were removed, as well as whole reads with more than 2 Ns after trimming. Short reads (less than 30 bp) were excluded from further analyses. Good quality paired-end reads which overlapped for at least 11 nucleotides were merged into a single consensus sequence. The software determines which of the two overlapping bases is of the highest quality and keeps it. The FastQC tool was used to verify the improvement of quality of the data produced and that adapter sequences have been completely removed. It is important to be sure that data is clean and suitable for downstream analyses.

5.2.6.4 Alignment and filtering

Mapping of the trimmed reads were made on the human reference genome hs37d5 with *aln* and *samse* commands of Burrows-Wheeler Aligner (BWA) v0.7.17 (*Li 2010*). Outputs are generated in the formats of .sai and .sam files, respectively. The *aln* algorithm is used for aDNA analyses because of better quality alignment of very short reads, which are typically less than 70 bp. For reducing the bias during the mapping step in aDNA analyses the edit distance which shows how dissimilar two sequences are by counting the minimum number of insertions, deletions or substitutions, was decreased to 0.01, which allows more mismatches, therefore reducing reference bias. To deal with ancient DNA post-mortem damage the maximum number of gap opens (which refers to the insertion or deletion of bases in the alignment), was increased to 2, and the seeding was disabled which makes the alignment more flexible (*Schubert 2012, Martiniano 2020*).

One of the produced types of files in .sam format was converted into a binary alignment map (.bam) file by using the software SAMtools v1.17. (*Li 2009*). This file facilitates the storage and computation, because it stores the same information as .sam but in a binary version. SAMtools also removes all low-quality reads (less than 20) and all unmapped reads, making the reads ready for the downstream steps of analyses.

5.2.6.5 Duplicates removal and indexing

Removal of duplicates is crucial for minimisation of biases introduced during PCR amplification, prevention of overestimation or false positivity in detection of variants. It ensures more accurate measuring of effective sequencing depth and confirms reliability of aDNA damage patterns.

The DeDup tool which is made for ultra-short DNA (*Peltzer 2016*) was used to remove duplicates. DeDup is specific for ultra-short DNA and removes reads with the same starting position at both 5' and 3' ends. After that .bam files were indexed with SAMtools and .bai files produced as outputs. This indexing gives efficient data retrieval for analysis. Indexing allows fast random access to different genomic regions, which in turn, ease the visualisation, enables faster downstream analyses for extraction of alignments of particular overlapping genomic regions.

5.2.6.6 Soft clipping

Variant calling and further analyses require a BAM file without post-mortem damage aDNA pattern. The half-UDG protocol for preparation of the libraries gives us mistakes in the last two nucleotides from both sides of a read, therefore the aim is to mask these positions. Hence, TrimBam from the bamUtil v. 1.0.15 software (*Jun 2015*) was used to soft clip the two terminal

bases at each end of the reads. Hence, soft clipping keeps the nucleotides in the reads, but labels them as not a part of the primary alignment. This maintains the overall length and integrity of the sequences while excluding potential confusion from further analysis.

5.2.6.7 Quality control

A quality control of the BAM files was made with Qualimap v.2.2.2-dev (*Okonechnikov 2016*). The tool provides summary statistics of the BAM file, such as number of mapped and unmapped reads, duplication rates, and overall alignment rates. Also, it shows depth of coverage across the genome, GC content of reads, the distribution of read lengths in the dataset. For the mapping statistics it provides mapping quality scores, which designate the estimates of the placement of reads on the reference genome.

5.2.6.8 Damage patterns analyses

The original BAM file (not soft clipped) can be used for estimation of the damage, error rate, and contamination and ancient DNA authentication, since it retains typical aDNA damage patterns in the terminal(s) bases of the reads. To make sure of the typical aDNA characteristics the software mapDamage v. 2.1.0 (*Jónsson 2013*) was used. Two types of input files are necessary for the software: a BAM file with the correct header and a FASTA file with included reference sequence for mapping the reads. To lower the running time, a downsampling option was used which randomly chooses a fraction of the reads, equivalent to 20% of the total.

The software mapDamage v.2.1.0 is designed for ancient DNA analyses. It is needed for assessing and quantifying the aDNA damage patterns and visualising them with graphic outputs. The software is able to identify and quantify C-to-T and G-to-A transitions, point higher damage rates at the ends of DNA fragments, which is typical for ancient DNA in particular, differentiate between ancient DNA and potential modern contaminants.

DNA fragments in ancient samples typically have higher damage rates at the ends compared to modern DNA. The mapDamage can pinpoint these end-specific damages, which are crucial for distinguishing ancient DNA from modern contaminants and length distribution of DNA fragments, showing the fragmentation patterns.

5.2.6.9 Error rate

The software Analysis of Next Generation Sequencing Data (ANGSD) v. 0.940 (*Korneliussen 2014*) was used for estimation of the error rate of the sequences reads. This method considers the excess of derived alleles in a sequenced individual compared to the derived alleles in an “error-free”

genome. The analysed genome and the “error-free” sequence should have the same number of derived alleles, while any additional polymorphisms found are the result of errors (*Orlando 2013*). The derived alleles are determined with respect to an outgroup sequence (we used the chimpanzee genome). As an “error-free” individual we considered a high-coverage genome from the 1000 Genomes Project (*Byrska-Bishop 2022*).

5.2.6.10 Sex estimation

The molecular sex of the individuals can be sometimes the only accessible estimation for some kind of badly preserved remains, and after all the previous steps of analysis of the sequenced data it is a highly reliable instrument of information, surely in case of good quality aDNA outputs (*White 2011*). Therefore, the molecular sex was identified by consideration of the number of alignments to chromosomes X and Y.

RY method

According to the method of (*Skoglund 2013*) a 95% normal approximation confidence interval (CI) is established using the RY approach and is computed as the number of Y chromosome alignments (ny) divided by the total number of sex chromosomal alignments (nx+ny). If the confidence interval lower bound exceeds 0.077 and the confidence interval upper bound is less than 0.016, the sample is categorised as male.

RX method

Another estimation of the sex is made according to (*Mittnik 2016*) which computes the Rx value, i.e. the average normalised ratio of the reads aligned to the X chromosome to the reads aligned to the 22 autosomes. In this method 95% normal approximation confidence interval (CI) is established as in RY. If a sample's CI upper bound is less than 0.60 it is deemed a male, while if its CI lower bound is greater than 0.80, the sample is considered a female.

5.2.6.11 Classification of mtDNA

As mtDNA is a circular molecule it has some difficulties with mapping it against rCRS. CircularMapper v.1.93.5 (*Peltzer 2016*) was therefore used to map reads to a linear reference sequence, but accounting for circularity. This tool elongates the linear representation of the genome by duplicating a small segment (500 nucleotides) from the beginning of the sequence and attaching it to the end. This allows to align correctly the reads which overlap the “end” of the linear genome, since these extra 500 nucleotides simulate the circular nature of the genome. After that, reads were again remapped to the standard rCRS with the same tool. Reads which did not pass the quality filter of a

mapping quality score of 30, unmapped reads, or duplicates were then discarded.

The newly produced BAM files were analysed with the ANGSD software to create a consensus FASTA file. For this output only reads with a minimum quality of 30 and positions with a minimum depth of 5 are considered.

The obtained FASTA file was run through the Haplogrep 3.2.1 software (*Schönherr 2023)*, which classified FASTA files assigning a haplotype and mitochondrial haplogroup for each sample.

5.2.6.12 Contamination estimation methods

Contamination of extracted DNA from ancient remains still has to be checked for the presence of any modern human contaminants. In case of detected contamination, it must be eliminated before going to further steps of analyses (*Racimo 2016*). Two different methods were used to evaluate the possible contamination. These methods are based on mtDNA investigation. One called ContamMix, another one Schmutzi. These methods are popular due to easier usage and more precise estimation for low-coverage DNA sequencing, because of the size and the amount of the copies of mtDNA per cell, allowing to use it as a proxy for nuclear DNA contamination. For each method used a BAM file generated at the very beginning (paragraph 5.2.6.4), which contains the whole picture of the damage pattern.

ContamMix

ContamMix v1.0-10 (*Fu 2013*) is a tool designed for estimation of contamination on aDNA sequences. The software is able to identify modern DNA contamination by comparison of the sequences from the aDNA sample to known modern reference genomes. ContamMix is working with low-coverage samples, yet with coverage that allows it to call the true endogenous mtDNA consensus sequence. It works with ancient samples as if they are sufficient and also it is considered as granted that contamination in the data cannot be higher than 50%. The consensus sequence was built from the alignment with ANGSD v.0.940 (*Korneliussen 2014*) with reads of mapping quality higher than 30 and base quality more than 20, and with parameters -doCounts 1 and -doFasta 2. All positions covered by less than three reads were discarded and considered as Ns.

After being converted to FASTQ, the original reads that mapped uniquely to the mtDNA reference sequence were remapped to this consensus. Multiple Alignment Fast Fourier Transform (MAFFT) v7.475 (*Katoh 2002, Katoh 2013*) was used to align the consensus sequence with a panel of 311 modern mtDNA genomes from around the world (*Green 2008*), which served as a source of probable contaminants. The multiple sequence alignment and the

newly created BAM files were used as inputs. To evaluate the degree of contamination, a Markov chain Monte Carlo (MCMC) framework was applied.

Schmutzi

Schmutzi is a software used for analysing contamination in mitochondrial aDNA data. Schmutzi utilises Bayesian statistical methods and likelihood calculations to find the differences between endogenous ancient sequences and contaminant modern sequences by considering measures of both deamination (considering that modern DNA is not deaminated) and fragment length distributions (*Renaud 2015*). The primary input is also BAM files. This estimate is used as an input for the next step, where contamination is evaluated using differences in SNPs between the endogenous mtDNA sequence and a database of potential contaminant modern mitochondrial genomes. The procedure is repeated until a solid estimate of the contamination rate is obtained and a single contaminating mitochondrial genome is found. In fact, the tool can output a consensus sequence of the found contaminant, for identification of the source. Important to consider that the protocol of library preparation that had been used with half-UDG treatment for removing typical aDNA damage can have a negative effect on this type of contamination estimates due to removal of deamination pattern.

HapCon

The software HapCon (*https://haproh.readthedocs.io/en/latest/hapCon.html*) works well with aDNA of modern *Homo Sapiens*, while it cannot be used for study of Neanderthals or Denisovans or other archaic hominins. Also, it has some difficulties with data from Sub-Saharan individuals, because of big amount of related ancestry, which is not covered in the 1000Genome dataset. Nevertheless, this method allows to obtain reliable results for WGS data with coverage as low as 0.02X on the male X chromosome or even 0.1X for SNP capture data (1240k). The software assumes that the haplotype of interest of X chromosome can be a mosaic copy from the modern reference sequences. Mismatches between the observed sequences and the copied haplotypes are modelled as either errors or contamination (*Huang 2022*). Finally, the software uses a Hidden Markov Model (HMM) to estimate contamination by maximum likelihood. We used default parameters together with 1000Genome reference panel for our male individual UKR-K4b4.

5.2.6.13 Pseudo-haploid variant calling

Given the low coverage nature of aDNA sequences, aDNA variant calling is usually performed in a “pseudo-haploid” manner. One allele is picked for each covered position by randomly sampling only one of the reads. The pseudo-haploid variant calling was done with ANGSD with option -doHaploCall 1 on

the 1240K SNP set. The output file comes in haplo.gz format, in which each row contains information about a SNP site, chromosome name, specific position of SNP and major allele in populations. The haplo.gz file must be converted into PLINK format files .tped and .tfam, which are required for further downstream analyses.

5.2.7 Kinship analysis

To find relatedness of the individuals in the datasets two different methods were used in this work. Ancient data were analysed with help of READv2 (Relatedness Estimation from Ancient DNA version 2) software. It uses pseudo-haploid input data (.tped and .tfam files) with the genome divided into 1 Mbp windows for estimation of pairwise mismatch rate per window. After that, it uses the genome-wide mean for classification of the relationship. The estimated pairwise mismatch rate on an unrelated pair of individuals from the same population is used to normalise the data and adjust for variations in background relatedness caused by SNP ascertainment and population diversification. Next, READ classifies pairs of individuals as identical/twins, first-degree relatives, second-degree relatives, third-degree relatives, or unrelated using this normalised pairwise mismatch rate (P0). After normalisation of the P0, this value is used to classify each pair of individuals as unrelated ($P0 \geq 0.90625$), second-degree relationship ($0.90625 < P0 \leq 0.8125$), first-degree relationship ($0.8125 < P0 \leq 0.625$) or identical twins or individual ($P0 < 0.625$) (*Alaçamlı 2024*).

Another method used for checking of the kinship was KIN tool (https://github.com/DivyaratanPopli/Kinship_Inference), which is used for kinship analysis of aDNA, designed to identify close relatives within a dataset by analysing genetic similarities. It measures the probability of two alleles of the same locus from different individuals to be identical by descent. As input files were used soft clipped .bam and .bai files with .bed file which contains information about the positions of SNPs to be analysed from 1240K panel.

Among ancient genomes from the territory of Ukraine were observed two Scythian individuals which were shown as identical or twins with the READv2 software, therefore we double checked it with the KIN tool. It was proved that the bones we obtained were from the same individual, also by this test, since the studied bones were coming from left and right temporal bones, we assumed and assured with colleagues in collaboration (P. Shydlovskyi, L. Samoilenko, C. Sharapova) that it could be destroyed cranium of the same individual. Therefore, for the further analysis the obtained BAM files were merged and analysed as one sample UKR-K1b1&2. The other individuals under study, according to a second-degree relatedness threshold, were found as not related.

5.2.8 Allele frequency analysis

All main steps for PCA were done as described in paragraph 5.1.2.3. The pseudo-haploid genotypes of ancient individuals were merged with a panel of modern populations genotyped with the Human Origin array (*Capodiferro 2021*) using PLINK v.1.9. This dataset was converted from PLINK into the EIGENSOFT format, using EIGENSOFT *convertf* tool (*Patterson 2012*).

Important modifications were done on the ancient DNA dataset. Smartpca was run using the option *lsqproject*, allowing the calculation of the principal components only of the modern populations and projecting on them the ancient individuals. For plotting the outputs, the *ggplot2* package of R 4.3.0 was used (*Wickham 2016).*

6. Results on modern individuals from present-day Ukraine (Donetsk oblast)

6.1 Mitochondrial variation in Donetsk oblast

After sequencing all 94 mtDNAs, we excluded three samples (UA11, UA72) due to very low-quality sequences or contamination (UA87). All these three samples belong to the group with Ukrainian TMA; two of them were from Donetsk oblast, and one from Mykolaiv oblast.

In the final dataset of 91 complete mitogenomes, the number of polymorphic sites (*S*) shows high variation among individuals (Table 6.1.1). Such genetic variation can be explained by extensive admixture and/or large effective population size. Haplotype diversity (*Hd*) evaluates the uniqueness of haplotypes within the population. The closer the value is to 1, the more genetic diversity is present in the population. The studied population displays very high diversity, in agreement with other neighbouring European populations (with sample sizes ≥ 80 and *Hd* = 0.999; Table 6.1.1). An interesting exception is Sardinia (*n* = 63 and *Hd* = 0.992), which is known for being an isolated population with a higher level of inbreeding. The high values of *Hd* of the Ukrainians, both published and studied in this research, suggest either a large stable population or recent expansions with new mutations differentiating haplotypes.

Table 6.1.1. *Comparison of genetic diversity indices among European populations*

Population	N samples	N haplotypes	S	Hd (std dev)	π (std dev)
Donetsk oblast	91	91	520	0.999 (0.00002)	0.0018 (0.0007)
Ukrainians *	144	141	na	0.999 (0.001)	0.0018 (0.0001)
Russians *	376	361	na	0.999 (0.000)	0.0018 (0.0001)
Czechs *	150	140	na	0.999 (0.001)	0.0017 (0.0001)
Slovaks *	139	133	na	0.999 (0.001)	0.0017 (0.0001)
Poles *	300	287	na	0.999 (0.0003)	0.0019 (0.0001)
Hungarians *	80	78	na	0.999 (0.002)	0.0018 (0.0001)
Sardinians *	63	50	na	0.992 (0.004)	0.0015 (0.0001)

S is Number of polymorphic sites; Hd is Haplotype diversity; π is Nucleotide diversity; the asterisk indicates data from (Malyarchuk 2023)

Nucleotide diversity (π) measures the average genetic differences per site when comparing pairs of sequences. The π value for the Donetsk oblast population is 0.0018 (Table 6.1.1), which is relatively low and indicates that, despite the large number of polymorphic sites, the overall genetic divergence among individuals is small. These data suggest that the Ukrainian population

may have passed through a bottleneck followed by rapid growth, as seen in other European populations. The even lower value in Sardinia confirms a closely related population with a recent common ancestry.

In conclusion, these data suggest that the current population of Donetsk oblast has experienced expansions and/or a complex admixture history (high haplotype diversity), but probably in recent times (low nucleotide diversity).

The maximum parsimony tree of the final dataset of 91 mitogenomes reveals extensive mtDNA variation within the region, with different branches and sub-branches corresponding to 80 haplogroups and sub-haplogroups (*Figure 6.1.1*).

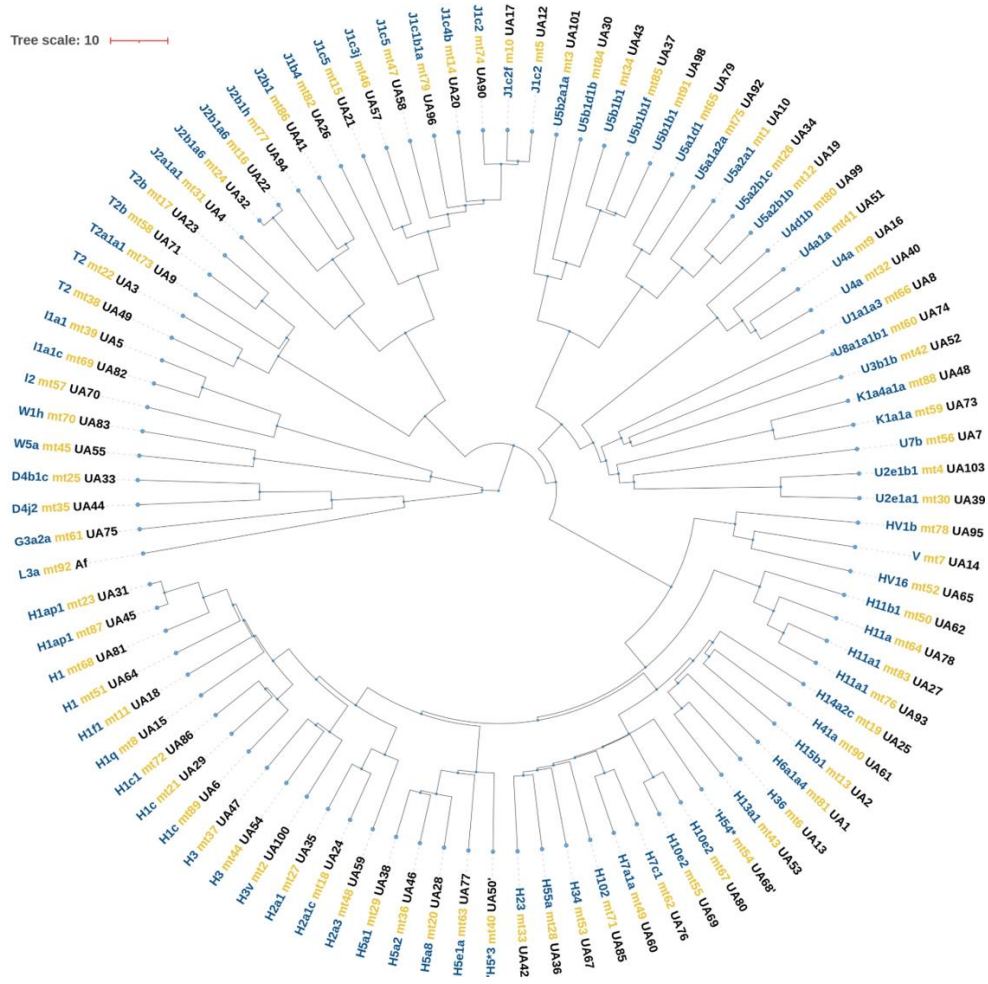


Figure 6.1.1. *Maximum parsimony tree of the entire Donetsk oblast mitogenome dataset with an African outgroup (Af) belonging to L3a*

The main branches of the tree represent 37 major haplogroups, most of which are of western Eurasian origin (96.7%; *Figure 6.1.2*). According to the literature (*Pshenichnov 2007, Malyarchuk 2023, Балановский 2012*), the most frequent haplogroup in this area is H. In our dataset, the most frequent haplogroup is also H (42.9%), which includes several sub-branches: H1-H3, H5-H7, H10, H11, H13-H15, H23, H34, H36, H41, H54, H55, H102. Other common haplogroups are U5 (11%) and J1 (10%). Haplogroups typical of western Eurasia were also found at very low frequencies, such as V and sub-branches of U, i.e. U4, U5a (U5a1 and U5a1a) and K. These data indicate similarities between the Ukrainian gene pool and other European populations. However, haplogroups of East Asia origin, D4 and G3, are also present at low frequencies (3.3% in total), consistent with previous studies on Ukraine where D, G and Z were reported at 1.6% (*Pshenichnov 2013*), 2.5% (*Mielnik-Sikorska 2013*), and 2.8% (*Malyarchuk 2023*) in total.

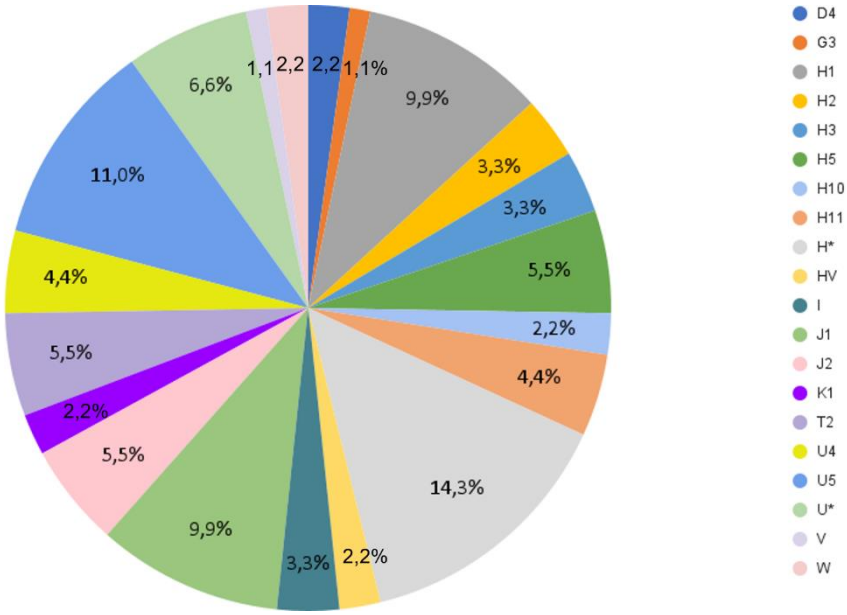


Figure 6.1.2. *Haplogroup frequencies in the present-day Donetsk oblast region*

These similarities with available mtDNA datasets on Ukrainian populations are also confirmed when considering the HVS1 sequences (nps 16024-16400) and coding-region genotyping data published in (*Балановский 2012*) (Table 6.1.2). The Ukrainians in that study were from western regions (Khmelnytska oblast, Lviv oblast, and Ivano-Frankivsk oblast), the central region (Cherkasy oblast), and Ukrainians from Belgorod oblast (present-day Russia). The Cossacks are represented by groups currently living in Russia (Kuban and Terek Cossacks), with some degree of descent from Ukrainian Cossacks.

Table 6.1.2. *Comparison of frequencies of major mtDNA haplogroups in East Slavic populations*

Mt Haplogroup	Ukrainians* (n= 610)	Cossacks* (n= 256)	Ukraine Donetsk (n= 91)
D	0.002	0.008	0.022
G	0.002	0.000	0.011
H	0.384	0.394	0.429
HV	0.035	0.015	0.022
I	0.027	0.038	0.033
J	0.086	0.106	0.153
T2	0.077	0.061	0.055
U4	0.052	0.023	0.044
U5	0.098	0.083	0.109
U*	0.057	0.069	0.065
K	0.049	0.068	0.022
V	0.050	0.015	0.011
W	0.026	0.008	0.022

**Ukrainians and Cossacks data taken from (Балановский 2012). Cossacks cohort represented by Kuban and Terek Cossacks*

In conclusion, our data are consistent with other mitochondrial datasets on Ukraine, already published on larger cohorts, thus confirming the goodness of the mitochondrial gene pool described in this thesis at the highest possible molecular resolution. The final picture describes a high level of mtDNA variation within the present-day Ukraine, testified by a variety of different haplogroups, mostly of western Eurasian origin.

To summarise the information embedded in the mtDNA haplogroups, we performed a principal component analysis (PCA). In *Figure 6.1.3*, we compared our high-coverage Ukrainian modern mitogenomes from Donetsk oblast with modern data reported in GenBank (*Nucleic Acids Research*, *2013*). This comparative dataset included 22 Ukrainians (labelled in green in the final plot) as well as 54 populations from Europe, western and southern Asia and North Africa.

The peculiar position of the Ukrainian GenBank dataset along the PC2, close to southern European and Middle Eastern populations, can be due to the low number of samples (only 22). In contrast, our Ukrainian dataset from Donetsk oblast is plotted near to European countries, while most of the Asian part of Russia is pushed farther away by the first and more significant principal component. This confirms a scenario already reported in the literature through analyses of hypervariable mtDNA sequences (*Pshenichnov 2007, Pshenichnov 2013*).

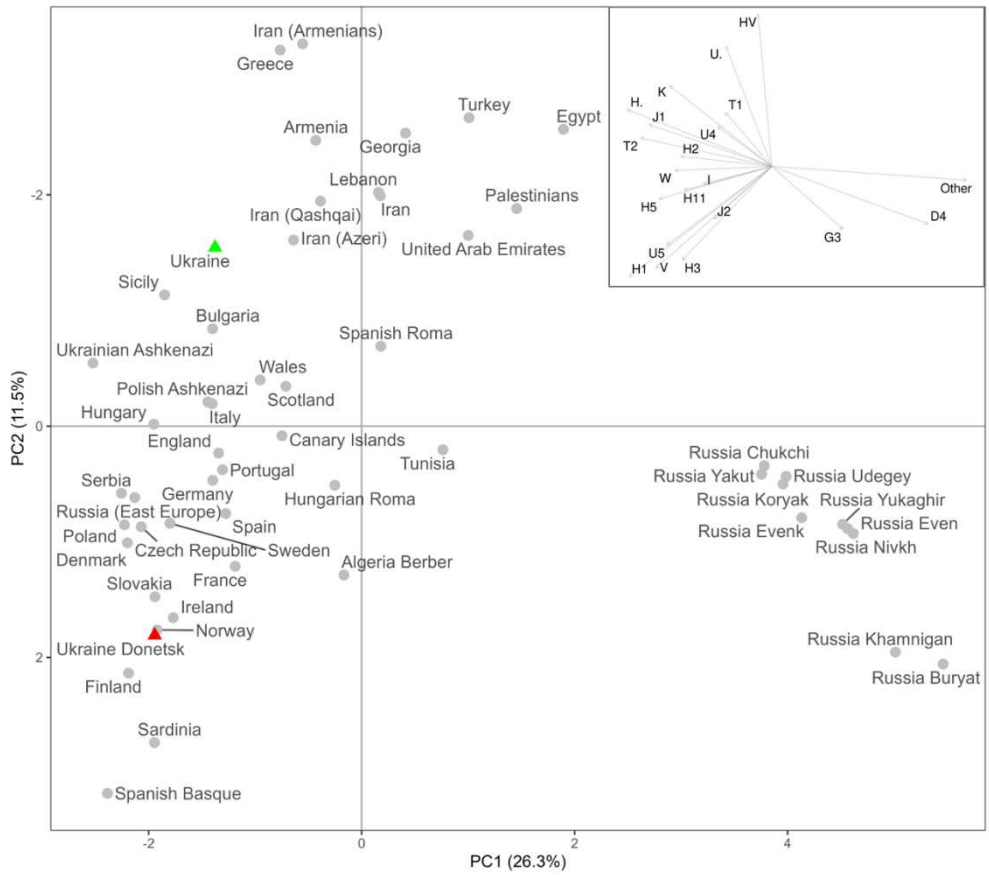


Figure 6.1.3. *PCA with complete mitogenomes from Europe, western and southern Asia, and North Africa. ▲- Ukrainian population from Donetsk oblast studied here, ▲- Ukrainian mitogenomes recorded in GeneBank*

6.1.1 Age estimates and demographic trends

Figure 6.1.1.1 shows age estimates for haplogroups that include at least three mitogenomes. These age estimates are consistent with those reported in the literature. The oldest estimate is for macrohaplogroup L3, which marks the out-of-Africa exit (*Torroni 2006*). It is dated between 60 and 70 kya, based on the entire dataset. Two of the most ancient western Eurasian (Palaeolithic) Hgs, U5 and U8, are both dated between 30 and 50 kya, while the ages of Hgs H1 and U5b1b1 (about 10-15 kya) point to the late glacial repopulation of western Eurasia (*Achilli 2004, 2005*). Notably, an H1 mitogenome has also been identified in an ancient Catacomb individual (see section 7.2.1). The more recent estimates of H5a and J1c2, less than 10 kya, could be related to movements of Neolithic farmers from the Near East (*Roostalu 2007*).

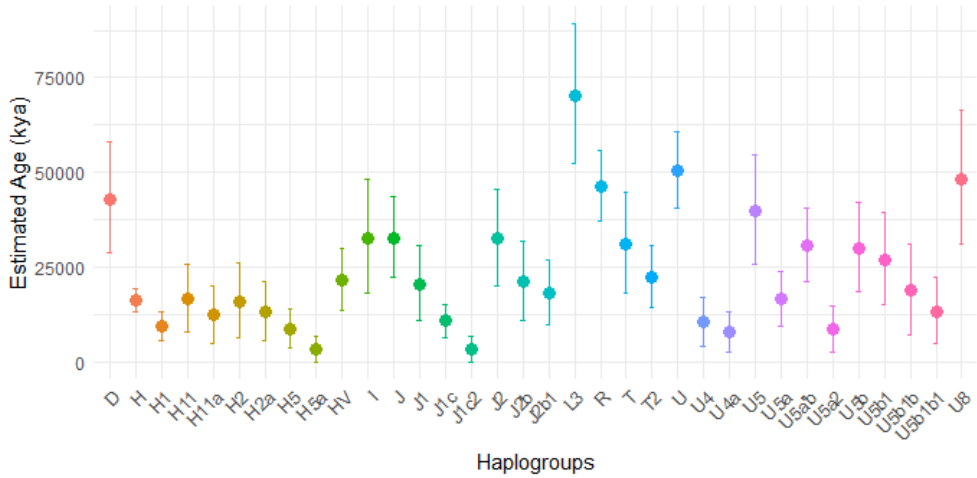


Figure 6.1.1.1. *Age estimates based on entire mitogenomes from Donetsk oblast plus two ancient mitogenomes (see section 7). Different colours represent different haplogroups*

The fact that the modern mitogenomes analysed here belong to haplogroups of different ages implies that the current mitochondrial gene pool of Donetsk oblast has been shaped by multiple migrations over time. This is further supported by the demographic trend depicted by the Bayesian Skyline Plot (BSP, *Figure 6.1.1.2*).

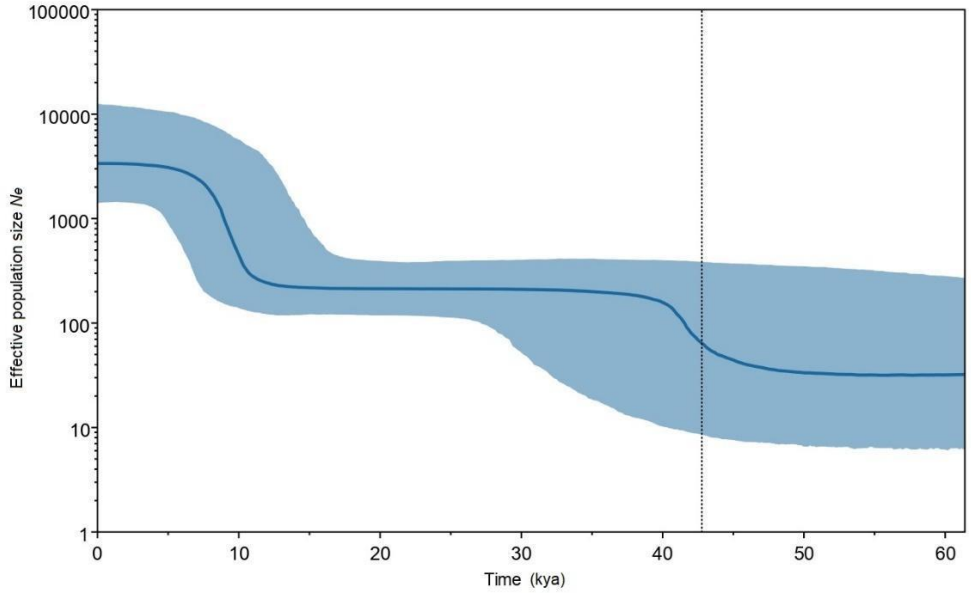


Figure 6.1.1.2. *Bayesian plot of the dynamics of the effective population size over time of the Donetsk oblast population. The plot shows the median values and the 95% confidence interval with the highest a posteriori probability*

An initial increase in the effective population size (*Ne*) started around 45 kya and lasted until about 38 kya, corresponding to the first peopling of Europe during the Paleolithic. A second increase is shown between 11 and 6 kya. This period is also known as the Neolithic Revolution associated with the spread of agricultural technology from the Near East.

In the literature (*Malyarchuk 2023*), a BSP for Ukrainian individuals and neighbouring populations from Russia reported a fluctuation in population size around 21 kya (corresponding to the Last Glacial Maximum). This may suggest the presence of human populations in eastern Europe (e.g., in glacial refugia). Additionally, it was found that the *Ne* of mixed cohorts from Ukraine and Russia has grown exponentially over the past 5,000 years beginning in the Bronze Age. This growth could reflect the spread of Kurgan culture bearers into northern and eastern Europe. In contrast, our plot, based on the Donetsk population, shows a plateau with a stable *Ne*, probably due to the small sample size.

6.2 Combining genealogical and mitochondrial data

After analysing the documents regarding the origin of the 91 mitogenomes, we identified 67 with TMAs from Ukraine and 24 from Russia. Among the probands whose DNA was sequenced and who had TMAs from Ukraine (n = 67), approximately 45% were from Donetsk oblast. To visualise the birthplaces of the Ukrainian TMAs, a map was created (*Figure 6.2.1*). The map illustrates the broad geographic distribution of the mtDNAs studied in this thesis. Given the comprehensive coverage across the country, this study not only describes a specific regional population but can also be extended to the national level.

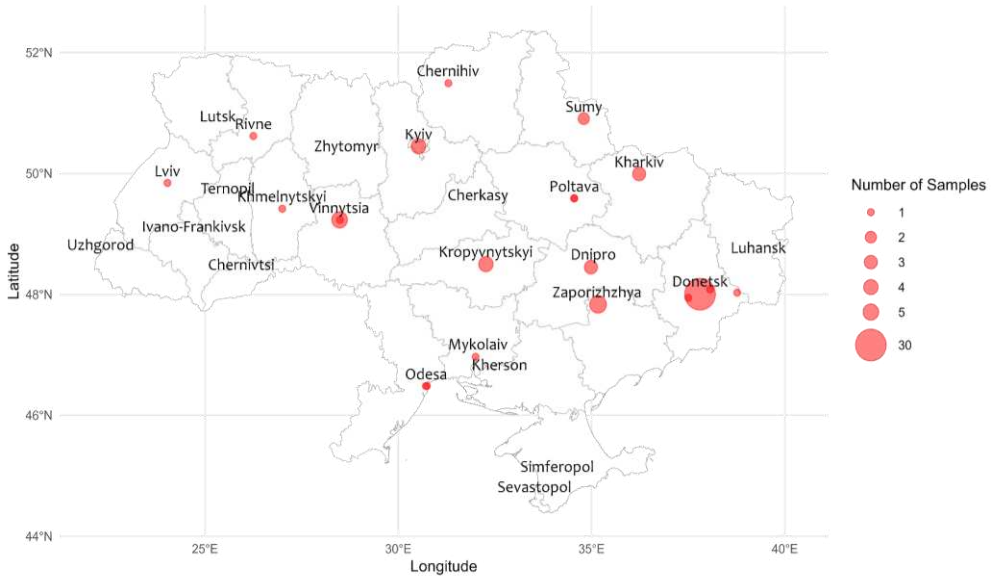


Figure 6.2.1. *Distribution of TMAs from Ukraine with birthplace localities, the size of the dots indicates the number of probands with the ancestry from a particular place or region. The dots represent exact localities when available; otherwise, they were placed in the administrative centre of each oblast*

When comparing the haplogroup frequencies between probands with TMAs from Ukraine and TMAs from Russia (*Figure 6.2.2*), no statistically significant

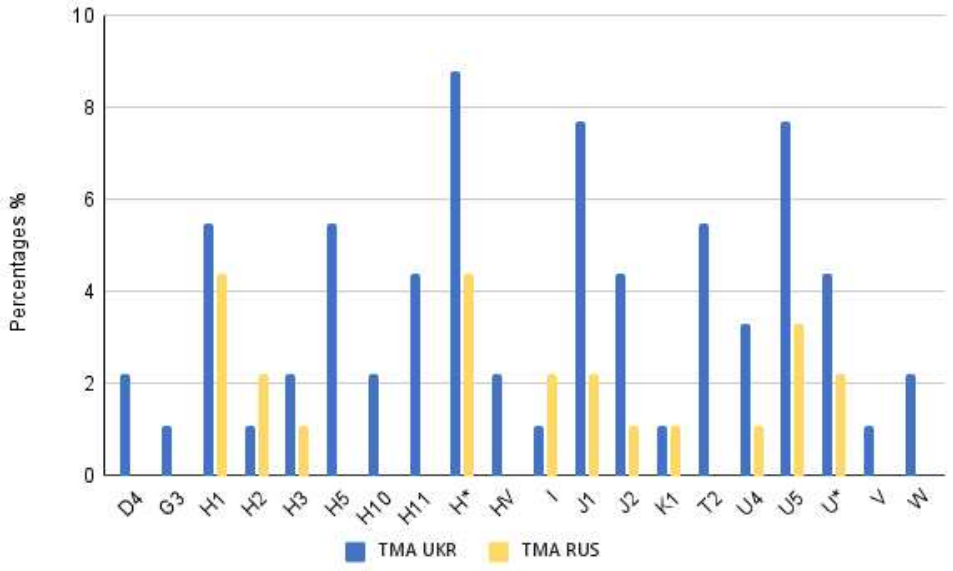


Figure 6.2.2. *Comparison of haplogroup frequencies in the two TMA subsets (Ukraine and Russia).* H* includes H6, H7, H13-H15, H23, H34, H36, H41, H54, H55, H108. U* includes U1-U3, U7, U8

difference was found ($p > 0.05$). However, the smaller number of mtDNAs with TMAs from Russia must be considered, which also accounts for the higher number of haplogroups in the Ukrainian dataset. In particular, two haplogroups of East Asian origin and seven from western Eurasia were found only in Ukraine. Taking into account the sample bias and the observed differences in the mtDNA gene pools of the two datasets, the possibility that Ukraine and Russia have different ancestral roots from a maternal perspective cannot be entirely ruled out.

6.3 Genome-wide profiles in Donetsk oblast

The DNA of 45 modern individuals from Donetsk oblast, whose grandparents were born in Ukraine, was genotyped using the Axiom Genome-Wide Human Origins 1 array. All 45 genotyped individuals passed the quality checks and were kept for further analysis.

6.3.1 Kinship

The kinship analysis performed with the KING software revealed no evidence of relationships up to the third degree, thus none of the genotypes were excluded from subsequent analyses.

6.3.2 PCA

The Principal Component Analysis of our 45 modern individuals from Donetsk oblast (with TMA and TPA from Ukraine) and 1469 modern individuals from Western Eurasian populations (*Mallick, Micco 2023, Mallick, Reich 2023, Reitsema 2022*) is shown in *figure 6.3.2.1*.

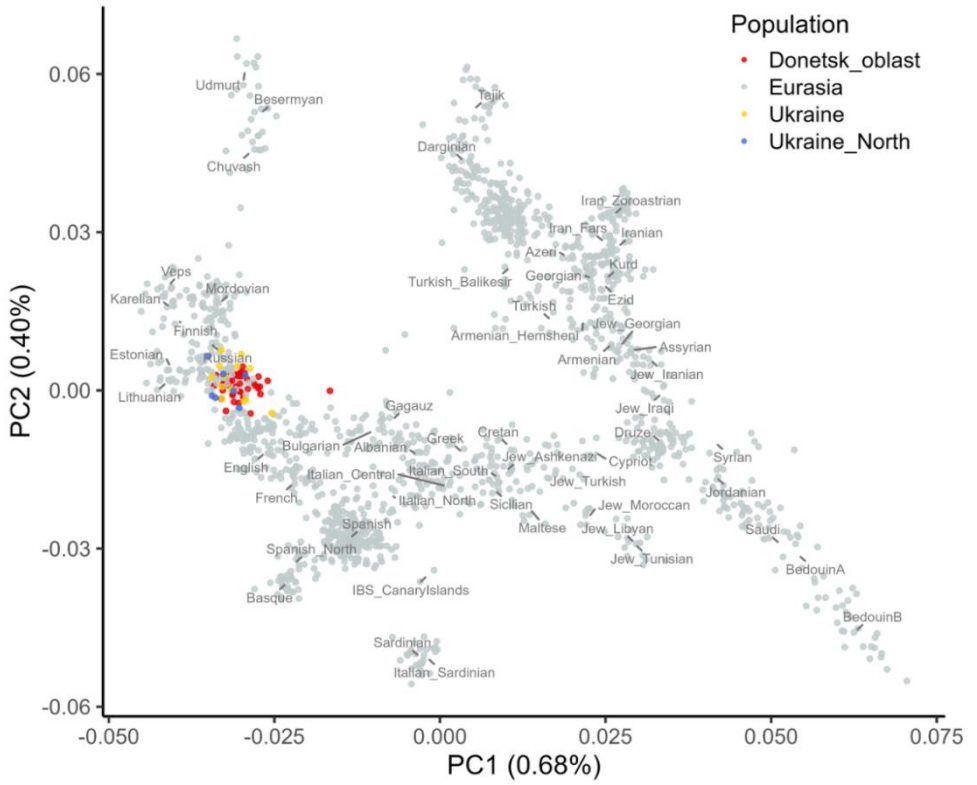


Figure 6.3.2.1. *PCA plot of 45 modern Donetsk oblast individuals (in red) with a comparative dataset of Western Eurasia populations (in grey). Other Ukrainian populations are highlighted in blue (northern Ukraine) and yellow (unspecified regions)*

The first principal component (PC1) accounts for the greatest variance among the populations and roughly separates the eastern countries from the western ones along the horizontal axis. The second principal component (PC2) divides northern European countries from southern ones, with a gradual transition. On the left side of the plot, we can observe a cline from northern European populations at the upper end toward Mediterranean ones. The cline on the right is characterised by Caucasian populations transitioning toward North African groups. Individuals from the Mediterranean region and Jewish communities fall between these two main genetic clines.

Ukrainian individuals from Donetsk oblast are plotted together with published Ukrainian genomes, with only one clear outlier (UA36, mtDNA haplogroup H55a) positioned closer to Balkan populations. Similar outliers with a genetic profile resembling those of southern Europe have been reported previously (*Oleksyk 2021*). The main Ukrainian cluster almost overlaps with an eastern Slavic cohort that includes Russian and Belarusian groups, and is close to northern European countries (e.g., England and Lithuania). The next closest

groups are Bulgarians and Gagauz, who are geographically located to the southwest of Ukraine.

6.3.3 Admixture

The same dataset used for the PCA was also used to perform an admixture analysis up to K 20. The smallest CV error was 0.424 at K= 4 (*Figure 6.3.3.1*).

The admixture analysis appears to confirm the genetic similarities between northern and eastern European groups. Compared to other modern populations of western Eurasia, Ukrainian individuals from Donetsk oblast have a similar pattern with respect to published Ukrainians, Belarusians, and Russians. In Lithuania the proportion of blue and purple components is also comparable, though without the contribution of the green ancestry.

Neolithic farmers are potential sources of the blue component, which is widely represented in all present-day European populations, such as France, Spain, Italy, northern and eastern Europe, and Ukraine. The purple component may derive from the steppe pastoralists of the Yamna culture, which connected Europe and Asia through widespread movements between 5 and 4.5 kya (*Lazaridis 2022*). The Yamna expansion also crossed the Caucasus around 4000 years ago, leaving genetic traces in the modern Caucasian populations. The green component, identified at low frequency in Ukrainians, may represent a complex ancestry derived from Neolithic Iran and Caucasus hunter-gatherers (*Clemente 2021).* Finally, the red component, more pronounced in a single Ukrainian individual from Donetsk, may confirm the influence of another Near Eastern ancestry, clearly observed in Levantine, Arabic, and Anatolian populations. These were probably other nomadic groups that spread pastoralism across the North Pontic-Caspian steppe (*Penske 2023*).

In summary, modern Ukrainians from Donetsk oblast have two main components derived from Neolithic farmers (blue), and Eneolithic steppe pastoralists (purple), but also experienced admixture events with Neolithic Iranian groups and Caucasian hunter-gatherers (green), as well as with Near Eastern farmers (red).

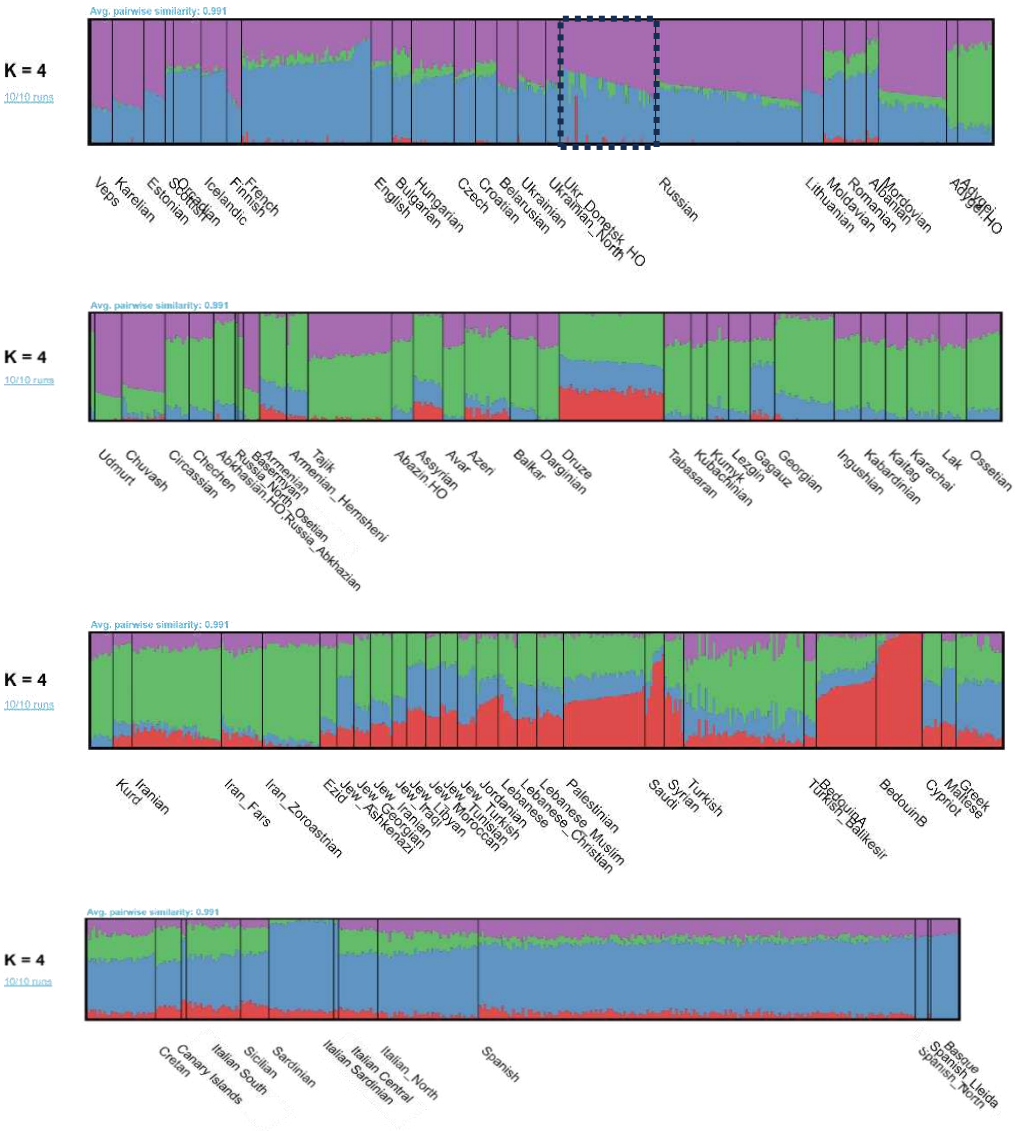


Figure 6.3.3.1 *Admixture plot including the population of Donetsk oblast at K=4*

In the reported barplot, each colour denotes an inferred ancestral population, and each thin vertical line represents an individual. The length of each vertical bar indicates the percentage of an individual's ancestry that comes from the inferred ancestral population associated with that colour

7. Results on ancient individuals from archaeological sites of present-day Ukraine

7.1 Anthropological context and radiocarbon dates

We selected bone or tooth fragments from seven individuals representing three different cultures and time periods (Yamna, Catacomb, and Scythian) based on anthropological assessments, which are summarised in table 7.1.1.

Table 7.1.1. *Anthropological information on individuals studied for this thesis*

Extraction ID	Skeletal element	Anthropol. sex	Anthropol. age	Bone Quality
UKR-b8	Molar I, 6 or 7	F	18-25	brown root, little cracks
	Molar III, 7			
UKR-K4b13	Temporal bone	M	25-30	Bad, porous
UKR-k21b-10A	Pre-Mo, upper	M	40+	Brown root, little cracks
	Temporal bone			Porous bone
UKR-k21b-10B	Temporal bone	F	35+	Outside porous, inside good
UKR-K4b4	Temporal bone	M	50+	Good
UKR-K1b1.1	Temporal bone	F	35+	Good
UKR-K1b1.2	Temporal bone	M	35+	Good

The results of radiocarbon dating confirmed the temporal and cultural contexts (Table 7.1.2). The Yamna culture existed in the territories of Ukraine from approximately 5,000 to 4,300 years before present (yr BP). The Catacomb culture spanned from 4,800 to 4,500 yr BP, with the Late Culture monuments dating back to 4,450–4,400 yr BP, and extreme dates ranging from 4,000 to 3,900 yr BP. Finally, the Scythian culture is dated from 2,700 to 2,200 yr BP in separate areas.

Table 7.1.2. *Radiocarbon dating results*

Culture	Sample code	^{14}C Age [yr BP]
Yamna	UKR-b8	4194±21
	UKR-K4b13	4165±21
Catacomb	UKR-k21b10A	*NA*
	UKR-k21b10B	3891±21
	UKR-K4b4	3928±21
Scythian	UKR-K1b1.1	2203±19
	UKR-K1b1.2	2179±19

NA - data not available

7.2 Ancient low-coverage genomes

After 44 extraction attempts, the sequencing results of 32 good quality libraries from seven ancient individuals are summarised in table 7.2.1.

Unfortunately, none of the sequencing efforts on the two Yamna individuals (UKR-b8 and UKR-K4b13) yielded sufficient endogenous reads for further analysis.

One (UKR-K4b4) of the three representatives of the Catacomb culture provided good sequencing results, which allowed us to obtain low-coverage genome data and to classify the mtDNA haplogroup. The endogenous content of this individual was estimated to range between 7.2% and 9.9% across three runs, with a merged average depth of 0.5X. The other two samples from a paired burial (UKR-k21b-10A and UKR-k21b-10b) yielded less than 0.03% endogenous DNA.

The aDNA extracted from the two Scythian individuals (UKR-K1b1.1. and UKR-K1b1.2) performed well, revealing an endogenous content between 3% and 33%, with a merged average depth of 0.6X. It is worth mentioning that anthropological and molecular sex determination differed for UKR-K1b1.2 (Tables 7.2.1 and 7.1.1).

In summary, we successfully obtained three low-coverage genomes (UKR-K1b1.1. UKR-K1b1.2 and UKR-K4b4) from the initial collection of seven individuals, achieving a 43% success rate.

Fragment length distributions, misincorporation patterns and error rates of these ancient genomes revealed typical patterns of ancient DNA, serving as a valuable tool to confirm the authenticity of the reads (see *Figure 7.2.1* for an example). The complexity curves suggest the potential to increase the average depth by performing deep sequencing on the same libraries.

Table 7.2.1. *Sequencing results on ancient individuals from Ukraine*

Culture	Sample ID	Read leng. avg	% duplicates	% endo gen. DNA	Avg depth (entire genome)	Sex	Avg depth MT	mtDNA Hg
Yamna	UKR-b8	47	12.28	0.02	<0.001	NA	0.01	NA
		53	10.21	0.03	<0.001	NA	0.01	NA
		47	12.28	0.02	<0.001	NA	0.01	NA
		53	10.21	0.03	<0.001	NA	0.01	NA
	UKR-K4b13	58	13.89	0.04	<0.001	NA	0.004	NA
		54	8.46	0.05	<0.001	NA	0.005	NA
		77	17.36	0.05	<0.001	NA	0.059	NA
		77	16.43	0.14	<0.001	NA	0.102	NA
		58	13.89	0.04	<0.001	NA	0.005	NA
		54	8.46	0.05	<0.001	NA	0.005	NA
Catacomb	UKR-k21b-10A	37	2.71	0.01	<0.001	NA	NA	NA
		37	3.11	0.02	<0.001	NA	0.01	NA
		39	2.29	0.01	<0.001	NA	NA	NA
		36	4.83	0.02	<0.001	NA	0.001	NA
		36	3.78	0.01	<0.001	NA	0.002	NA
		37	3.97	0.01	<0.001	NA	0.002	NA
		36	3.37	0.02	<0.001	NA	NA	NA
		54	14.97	0.01	<0.001	NA	0.003	NA
		54	13.33	0.26	0.001	NA	0.139	NA
		46	12.70	0.02	<0.001	NA	0.004	NA
	UKR-k21b-10B	35	13.61	0.01	<0.001	NA	0.004	NA
		36	15.94	0.03	<0.001	NA	0.006	NA
		38	14.99	0.01	<0.001	NA	NA	NA
		40	12.62	0.01	<0.001	NA	0.003	NA
	UKR-K4b4	50	16.59	9.86	0.025	M	2.89	H1b
		53	24.28	7.54	0.263	M	28.81	H1b
		53	23.47	7.20	0.354	M	38.02	H1b
Scythian	UKR-K1b1.1	54	16.7	20.2	0.172	F	27.73	D5a2a2
		54	1.09	8.56	0.012	F	1.62	D5a2
	UKR-K1b1.2	55	18.62	33.10	0.337	F	30.40	D5a2a2
		52	10.95	3.86	0.031	F	3.87	D5a2a2
		68	1.49	3.05	0.004	F	0.54	NA

Avg – average; Mt – mitochondrial; leng. – length; NA – not available

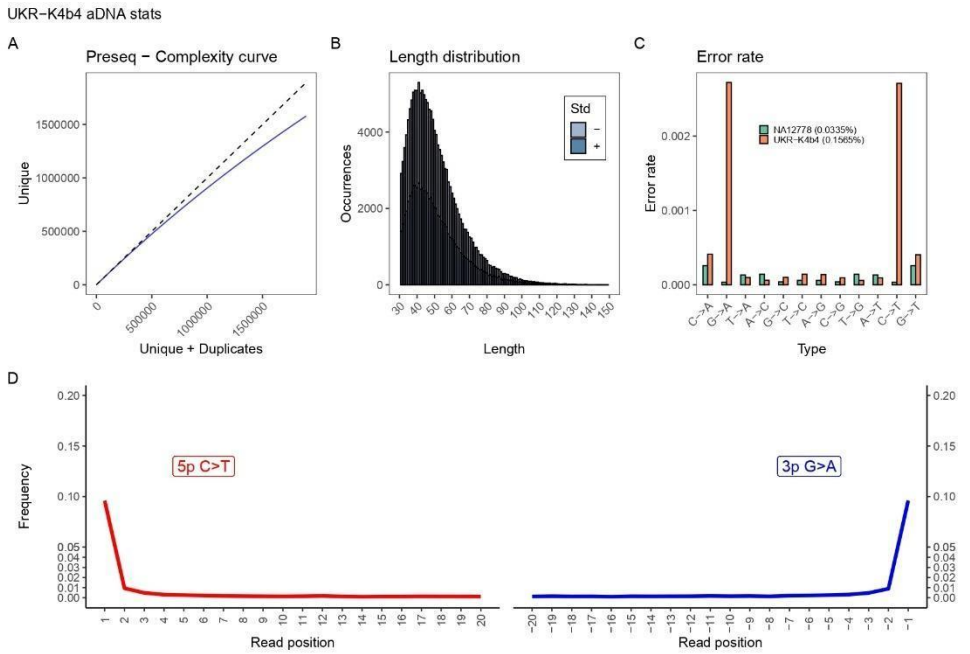


Figure 7.2.1. *Sequence analysis of the UKR-K4b4 reads. (A) complexity of the library, (B) the length fragment distribution with a peak around 50-70 bp, (C) error rates compared to a modern DNA individual, and (D) the frequency of transition (C to T and G to A) at the end of the reads*

Due to the low coverage, three methods were employed to estimate the level of contamination and confirm the results. ContamMix provides the probability of the authenticity of the endogenous sequence, with higher values being preferable, ideally approaching 1. In contrast, Schmutzi and hapCon estimate contamination, where lower values are preferable, ideally close to 0 with an acceptable threshold up to 3%. All three methods provide a confidence range for the contamination estimates, shown in parentheses (Table 7.2.2).

Table 7.2.2. *Contamination estimates with confidence intervals for UKR-K1b1.2. UKR-K1b1.1 and UKR-K4b4*

Culture	Sample ID	Sex	ContamMix	Schmutzi	HapCon
Scythian	UKR-K1b1.1-Co	F	0.994 (0.982-0.999)	0.01 (0-0.02)	*NA*
	UKR-K1b1.2-Co	F	0.982 (0.968-0.991)	0.02 (0.01-0.03)	*NA*
Catacomb	UKR-K4b4	M	0.962 (0.811-0.992)	0.99 (0.98-0.99)	0.000 (-0.007-0.007)

NA means undetermined

For samples UKR-K1b1.1 and UKR-K1b1.2, hapCon could not be applied, because it relies on haploid chromosomes, specifically the X chromosome in

males, and these individuals were female. The high contamination level detected by the Schmutzi method in sample UKR-K4b4 is likely due to the low coverage depth. In fact, the other two methods indicated no contamination. Therefore, all three sequences are considered to be uncontaminated.

7.2.1 Same ancient individual revealed by kinship analysis

An interesting outcome emerged from the kinship analysis: READv2 revealed that the two Scythian genomes belong to the same female individual. To increase the robustness of the results, additional Scythian genomes from a published study (*Gnecchi-Ruscone 2021*) were included in the analysis. *Figure 7.2.1.1* confirms the initial results for the two genomes under study.

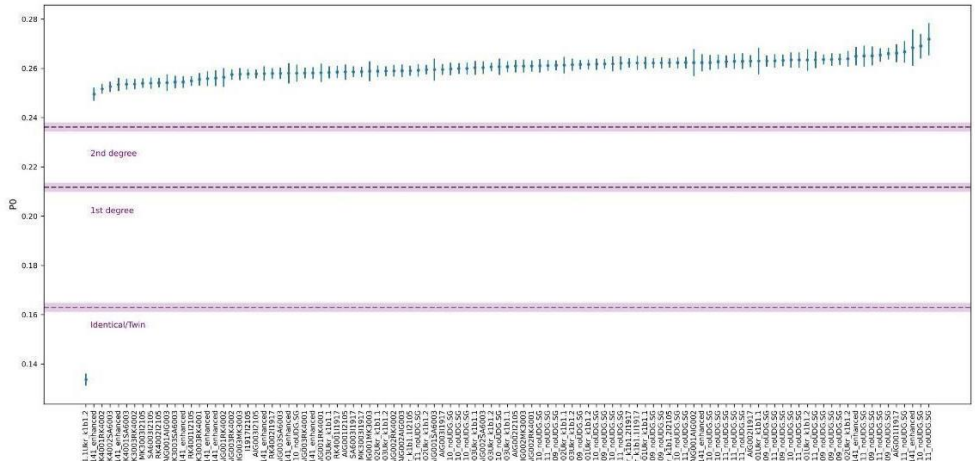


Figure 7.2.1.1. *Kinship results of UKR-K1b1.1 and UKR-K1b1.2 samples with the READv2 software*

Additional confirmation of the relatedness of these two individuals was obtained with KIN analyses, which were performed three times with different sample sizes. The same data set was used for both READv2 and KIN.

All results supported the initial hypothesis that these two individuals were identical (Table 7.2.1.1). Additionally, we noted that the two original ancient bones were left and right temporal bones (*Figure 7.2.1.2*), consistent with the possibility that they were from the same individual.

Table 7.2.1.1. *KIN results for UKR-K1b1.1 and UKR-K1b1.2-Co samples.*

KIN run№	Pair	Related-ness	Second Guess	Log Likelih Ratio	k0	k1	k2	IBD Len.	IBD Num
1	UKR-K1b1.1 – UKR-K1b1.2	Identical	Parent-Child	1.334	0	0	1	0	1
2	UKR-K1b1.1 – UKR-K1b1.2	Identical	Siblings	8.008	0	0	1	0	1
3	UKR-K1b1.1 – UKR-K1b1.2	Identical	Siblings	46.301	0	0	1	0	1

IBD – Identical by Descent, related to the number of DNA segments shared between individuals; k0-k3 – probability that two individuals share 0, 1 or 2 alleles IBD at a given locus respectively

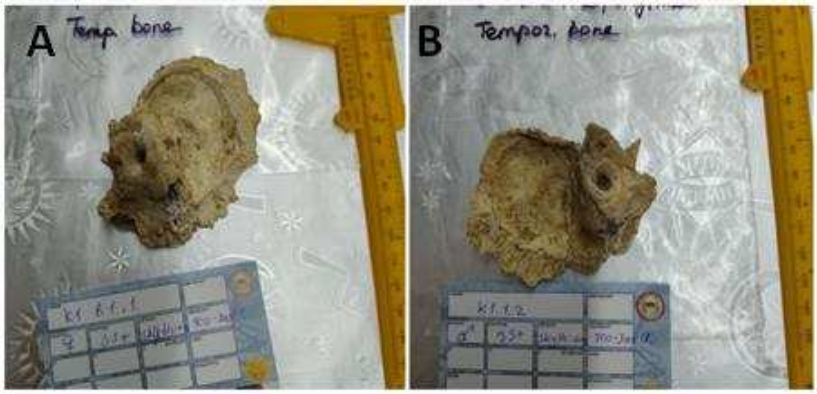


Figure 7.2.1.2. *Left (A) and right (B) temporal bones corresponding to samples UKR-K1b1.1 and UKR-K1b1.2*

We merged the original reads from UKR-K1b1.1 and UKR-K1b1.2 into an ancient genome named UKR-K1b1&2. All quality controls, including contamination tests, were repeated on this merged genome, confirming the typical ancient DNA pattern (*Figure 7.2.1.3*).

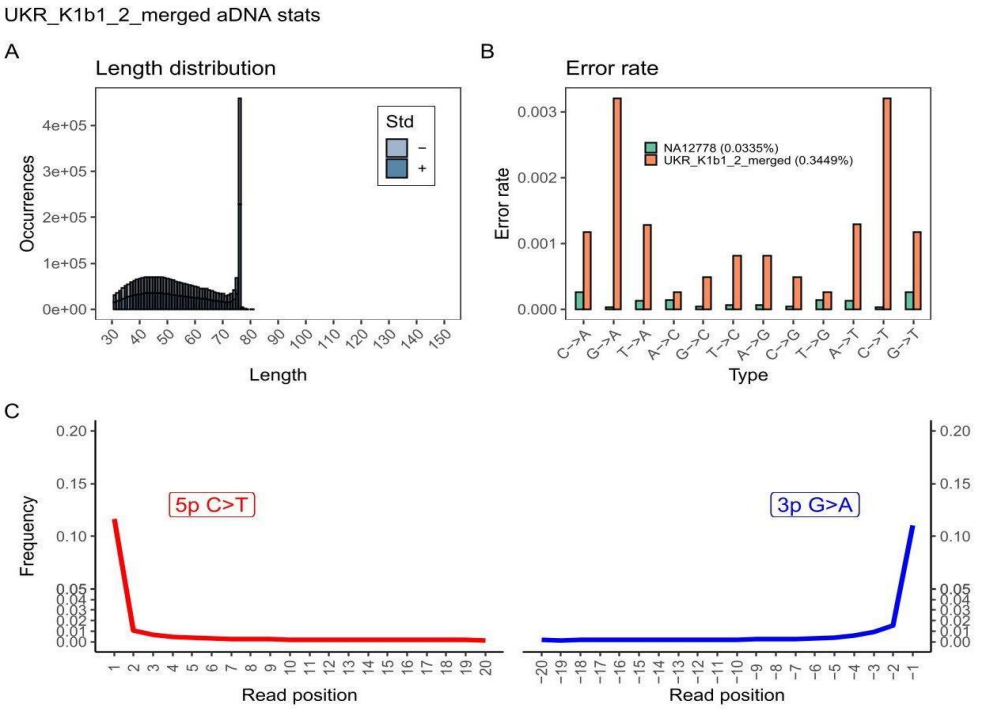


Figure 7.2.1.3. *Sequence analysis of the UKR-K1b1&2 reads. (A) the length fragment distribution with a peak around 50-70 bp, (B) error rates compared to a modern DNA individual, and (C) the frequency of transition (C to T and G to A) at the end of the reads*

Finally, by merging together the sequencing reads obtained from UKR-K4b4 and UKR K1b1&2 we were able to obtain two ancient genomes with a depth of coverage higher than 0.5X (Table 7.2.1.2).

Table 7.2.1.2. *Results on merged genomes of ancient individuals for UKR-K1b1&2 and UKR-K4b4*

Sample ID	Sex	Read length avg	% endogen. DNA	Avg depth (entire genome)
UKR-K1b1&2	F	55	27.17	0.51
UKR-K4b4	M	53	7.42	0.64

Avg – average

7.2.2 Ancient mitochondrial haplogroups

Thanks to the increased coverage, we were able to classify the two ancient mitogenomes into specific haplogroups with high confidence (Table 7.2.2.1). The individual UKR-K4b4, associated with the Catacomb culture, carries mitochondrial haplogroup H1b, commonly found in eastern Europe and frequently observed in Slavic populations, suggesting a matrilineal genetic continuity in the region. On the other hand, the Scythian sample UKR-K1b1-2 exhibits haplogroup D5a2a2, a subclade of the typical East Asian haplogroup D, which has been detected at low frequencies in modern Ukrainian populations (see Chapter 6.1). The identification of a typical Western Eurasian haplogroup (H1) alongside an East Asian one (D), suggests a complex pattern of migration that has shaped the gene pool of this area over time.

Table 7.2.2.1. *Haplogroup classification of individuals UKR-K4b4 and UKR-K1b1&2*

Individual	Haplogroup	Quality	mtDNA depth of coverage	Covered sites
UKR-K1b1&2	D5a2a2	0.981	58.126	18292
UKR-K4b4	H1b	0.953	69.725	23751

The ancient samples from this study were incorporated into the maximum parsimony tree of modern mitogenomes from Donetsk oblast (*Figure 7.2.2.1*). The results show that the ancient mitogenomes are embedded within modern mitochondrial branches. In particular, D5a2a2 is a sister clade of the modern D4 branches, both of East Asian origin. Haplogroup C, another subclade of the East Asian macrohaplogroup M, has been identified in the territory of present-day Ukraine during the Neolithic (*Juras 2017*). Other study on Scythians found on the territory of Ukraine (*Nikitin 2012*) also reported haplogroup D among the studied individuals. The ancient H1b, together with nine others modern mitogenomes, clusters within the H1 branch, which dates back over 10 kya (see *Figure 6.1.1.1*) suggesting the continuity of Ukrainian mitochondrial lineages over thousands of years.

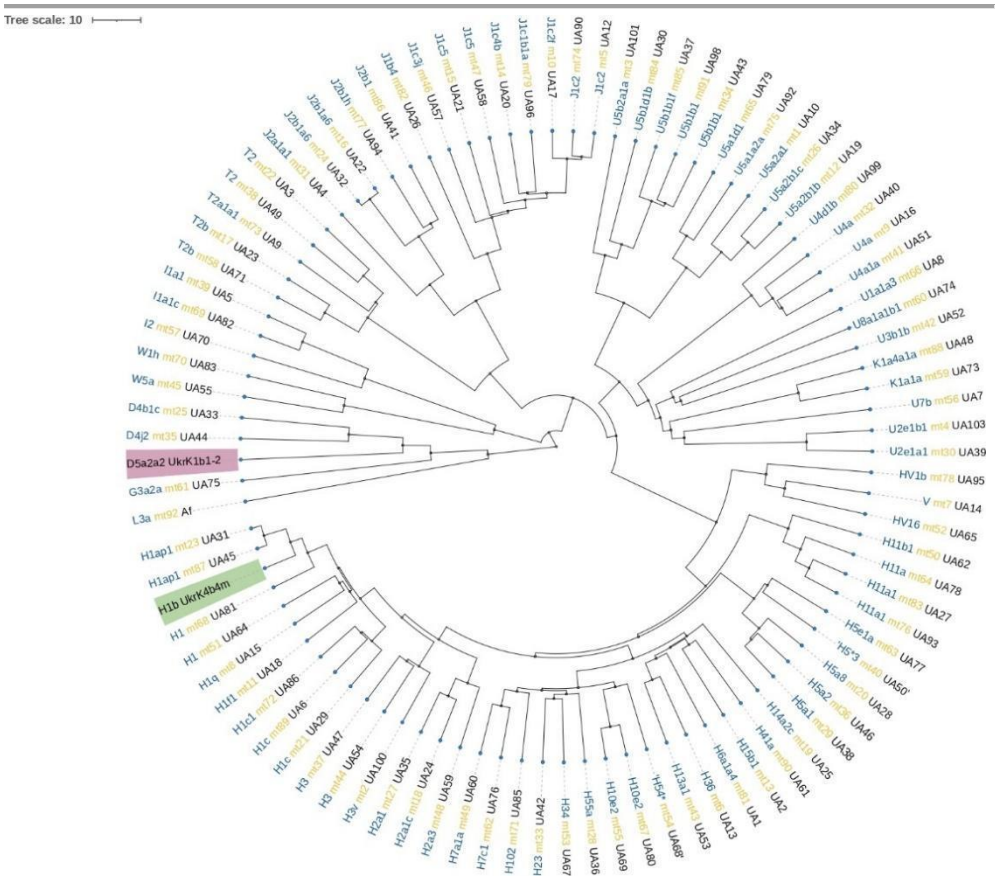


Figure 7.2.2.1. *Maximum parsimony tree of the entire Donetsk oblast mitogenome dataset plus two ancient mitogenomes (UKR-K1b1&2 in purple and UKR-K4b4 in green) from archaeological sites of present-day Ukraine. Af indicates an African outgroup belonging to L3a*

7.2.3 Genetic relationships with other populations

To assess the relatedness of ancient individuals to modern populations and compare the obtained ancient genomes with other published ones, PCA was performed by "projecting" the ancient genomes onto modern data. The same dataset used in Chapter 6.3.2 was applied for these analyses. In the PCA, the first principal component (PC1) roughly separates eastern and western countries along the horizontal axis (*Figure 7.2.3.1*). The second principal component (PC2) mostly distinguishes northern from southern populations.

The Catacomb individual UKR-K4b4 plots near other published Catacomb individuals (blue star on the plot) and occupies a position between Northern European populations and Turkic ethnic groups such as Besermyan, Udmurt, and Chuvash, located in the Volga federal district of Russia, west of the Ural Mountains (*Figure 7.2.3.1*). This positioning is consistent with the spread of Kurgan cultures. The nearest ancient genomes to UKR-K4b4 are from the Yamna and Corded Ware cultures, which share overlapping territories, cultural practices, and, to some extent, time periods.

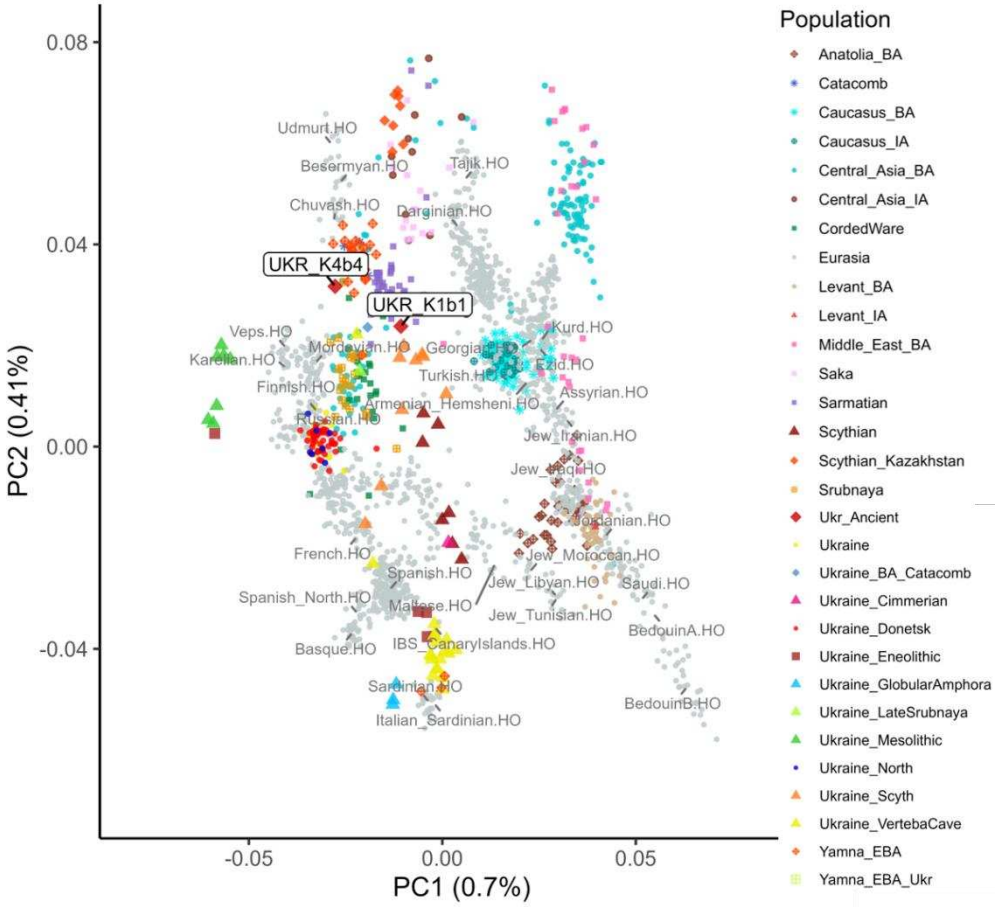


Figure 7.2.3.1. *PCA plot of 45 modern Donetsk oblast individuals (in red) and two ancients (UKR-K1b1&2 and UKR-K4b4) with a comparative dataset (Mallick 2023; Reitsema et al., 2022) of Western Eurasia populations (in grey). The modern Ukrainian populations are highlighted: North Ukraine in blue, not specified Ukraine in yellow*

The Scythian genome UKR-K1b1&2 occupies an intermediate position between Scythians from Kazakhstan and those found in Ukraine, situated along a continuum from Western Eurasia to the East (*Figure 7.2.3.1*). Scythians reported from Ukraine and neighbouring regions, particularly Russia, take a central position between Northern and Eastern European countries and the cline from the Caucasus and the Middle East. This positioning likely reflects significant admixture resulting from their nomadic lifestyle and expansions across the Eurasian steppes. The Scythian civilization stretched from present-day Ukraine, including the Crimea, to southern Russia and parts of Kazakhstan, encompassing lands of Anatolia and regions along the Danube river. The closest ancient populations to the studied individual are the Scythians found in Ukraine and the Sarmatians, which were one of the groups that replaced the Scythian civilization around the 3rd century BCE.

8. Discussion

The aim of this thesis was to investigate the genetic variability of the Ukrainian population to reconstruct its origins and history on the map of Europe, through separate and combined analyses of genomic data from the modern population of Donetsk oblast and past individuals who lived in the territory of present-day Ukraine.

The mitochondrial DNA analysis presented in this study provides crucial insights into matrilineal history. First, using genealogical data, we grouped individuals by their TMA birthplace and found that one-third were from western Russia, which is consistent with the recent history of Donetsk oblast. Extensive mitochondrial DNA variation was revealed in the 91 mitogenomes obtained via Illumina sequencing. These genomes show high haplotype diversity (*Hd:* 0.999) and depict a complex phylogenetic history. The maximum parsimony tree identified 37 major branches with 80 haplogroups and sub-haplogroups, predominantly of Western Eurasian origin (96.7%). Haplogroups D4 and G3, of East Asian origin, made up 3.3% of the total, aligning with previously published studies. A differential distribution of mitochondrial haplogroups was observed between individuals from Ukraine and Russia, suggesting different female ancestral roots, although the differences were not statistically significant.

The age estimates obtained on these haplogroups suggest that Donetsk's mitochondrial gene pool has been shaped by multiple migrations, confirming many events in Western Eurasian history. Macrohaplogroup L3 (60–70 kya) was involved in the out-of-Africa exit; Paleolithic haplogroups U5 and U8 (30–50 kya) indicate the initial settlement of hunter-gatherers in Western Eurasia; H1 and U5b1b1 (~10–15 kya) correspond to the late glacial repopulation of western Eurasia. Finally, more recent haplogroups H5a and J1c2 (<10 kya) align with Neolithic migrations from the Near East.

The overall matrilineal demography depicted by the Bayesian Skyline Plot (BSP) reveals two demographic expansions: one around 45–38 kya, corresponding to the first peopling of Europe, and another between 11–6 kya, coinciding with the Neolithic Revolution. A similar pattern has already been reported for the Ukrainian and Russian mitogenomes (*Malyarchuk 2023*), but with two additional fluctuations: one at about 20 kya, roughly at the end of the Last Glacial Maximum; the other over the last 5,000 years, probably linked to the spread of the Kurgan cultures. These fluctuations were not visible in the Donetsk oblast plot, likely due to the smaller sample size. However, this recent expansion could be suggested by the low nucleotide diversity (π*:* 0.0018) despite the presence of many polymorphic sites, which together could indicate a recent bottleneck followed by rapid growth. Biparental

information was obtained from the genome-wide analysis (~629,000 SNPs) of 45 modern individuals from Donetsk oblast (with TMA and TPA from Ukraine).

In the genomic landscape of Western Eurasia depicted by Principal Component Analysis (PCA), the Donetsk group clusters with other published Ukrainian genomes and overlaps with an Eastern Slavic cohort, together with Russian and Belarusian populations, positioned near Northern European populations. The closest non-Slavic groups are Bulgarians and Gagauz from the southwestern border of Ukraine. Admixture analysis confirms genetic similarities between northern and eastern European groups. Ukrainians from Donetsk oblast show patterns comparable to published Ukrainians, Belarusians, and Russians, with similarities to Lithuanians. Among the four genetic components identified, two major ones can be linked to Neolithic farmers and Bronze Age steppe pastoralists of the Yamna culture, respectively. Two smaller contributions could be related to a complex ancestry of Iranian farmers and Caucasian hunter-gatherers, as well as connections with the Levant, Arabia, and Anatolia.

The past of present-day Ukraine was also explored through the analysis of seven individuals from different cultures and time periods: Yamna, Catacomb, and Scythian. We were able to obtain mitochondrial and low-coverage genomic data by shotgun sequencing of only three samples. However, kinship analyses revealed that two petrous bones, from the left and right temporal bones, came from the same Scythian individual. This finding highlights the need for rigorous quality control in ancient DNA analysis and underscores the importance of incorporating anthropological and molecular data to confirm identities. For this merged Scythian genome, as well as for a Catacomb one, we were able to obtain a depth of coverage higher than 0.5X. One of the most significant observations regarding these low-coverage genomes concerns the identification of haplogroups of different origins. Haplogroup H1b, associated with the Catacomb culture, testifies to a link with Western Europe and suggests genetic continuity in the region, given its presence in modern Ukrainians since the Late Pleistocene/Early Holocene. On the other hand, the detection of haplogroup D5a2a2 in the Scythian sample points to fluxes from East Asia, probably reflecting the dynamic interactions among steppe populations since the Late Bronze Age. The presence of another D sub-haplogroup in contemporary Ukrainians further highlights the role of ancient nomadic movements in shaping the Ukrainian gene pool. These findings emphasise both continuity and diversity within the mitochondrial genetic pool, which has been shaped by multiple matrilineal waves over millennia.

When these two ancient genomes are projected onto the Western Eurasian landscape depicted by the PCA, the Catacomb individual consistently clusters with other published Catacomb individuals as well as with those

available for the Yamna and Corded Ware cultures, which shared similar territories. This “steppe” group lies between Northern European and Turkic ethnic groups, supporting existing theories about the origins of the Kurgan cultures. The closest ancient populations to our Scythian individual include the Sarmatians, who succeeded the Scythians around the 3rd century BCE. They fall between other Scythian groups from Kazakhstan and Ukraine and lie along a continuum from Western Eurasia to the east, reflecting the extensive nomadic lifestyle of the Scythians.

In conclusion, mitochondrial and genome-wide analyses of present-day individuals have revealed that the genetic landscape of Donetsk oblast results from complex migratory dynamics that have also shaped most of other Eurasian regions. The comparison of ancient Scythian and Catacomb genomes further supports this scenario, but also demonstrates long-term genetic continuity in the region, even if it was traversed by nomadic groups.

This dissertation adds another important piece of information to the complex genetic landscape of Eastern Europe. Future studies with larger sample sizes and deeper sequencing could further refine these insights and provide a more nuanced understanding of the genetic history of Europe.

9. Additional project: Medieval individuals from "Sant'Anna di Sopramonte" (Trentino, Italy)

Eleven Medieval burials were recently discovered in Sopramonte near Trento (*Figure 9.1*). Seven graves were found in 2019 and four more in 2021. Historical documents indicate that the cemetery was located in the immediate vicinity of the historic monastery of Sant'Anna. This has also been confirmed by detailed stratigraphic studies of the area. In "Tomb 1" was found a particularly interesting skeleton of a high-ranking person, buried in clothes embroidered with gold thread.

Figure 9.1. *Excavation site and human remains of "Sant'Anna di Sopramonte" (credits to Fabio Peterlongo, S.I.E. S.p.A. Società Iniziative Editoriali)*

The historical centre of Sant'Anna is located two kilometres above the town of Sopramonte and includes the church dedicated to the saint of the same name and the nearby "Casa del Preposto". In this place, some wall structures, about one metre deep, have been found. They can be traced back to a monastery that housed a community of nuns and monks between the thirteenth and the end of the fifteenth century. Historical sources present it as a significant political and religious entity of its time. The documents issued by Pope Urban IV in 1263 granted monks and nuns the privilege of being buried in the local cemetery.

The excavation was carried out by Christian Fogarolli and Mattia Segata under the coordination of the Superintendency for Cultural Heritage of the Autonomous Province of Trento (*Fabio Peterlongo, S.I.E. S.p.A. Società Iniziative Editoriali*).

Four molars were sent from the archaeological site of Sant'Anna Sopramonte to the University of Pavia for ancient DNA analysis, thanks to collaboration with Professor Luca Pagani from the University of Padua. They were dated between 1100 and 1400 CE based on the archaeological context and historical knowledge. A more accurate radiocarbon dating is still in progress.

Molars are among the best parts of the human body to preserve ancient DNA (*Dabney 2013*). The extraction methods used for these four teeth were different. One method requires cutting and powdering the whole tooth root (WTR method). The minimally destructive extraction (MDE) method aims to obtain ancient DNA from the dental cementum that is located on the surface of the teeth roots, see chapters 5.2.3.3 and 5.2.3.4.

All extracted DNA samples were quantified using a Qubit Fluorometer and evaluated through capillary electrophoresis (*Figure 9.2*).

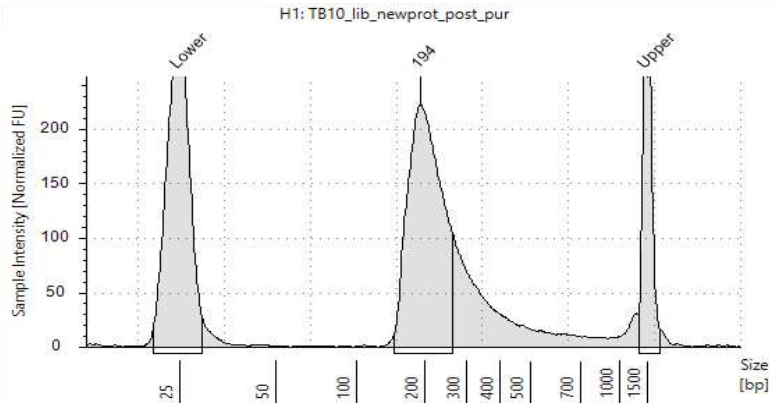


Figure 9.2. *Analysis of the Tb10 DNA sample with the Tape Station 4150 (Agilent, USA) to check the quality and size of the extracted DNA fragment (after purification).*

Half-UDG libraries (*Meyer 2010*) were sequenced on the Illumina NextSeq-500 platform at the Genomic and Post-Genomic Unit, IRCCS Mondino Foundation in Pavia. The raw reads (2x75bp pair-ends) were validated by estimating standard ancient DNA parameters, such as fragmentation, error rate and misincorporation pattern (*Figure 9.3*).

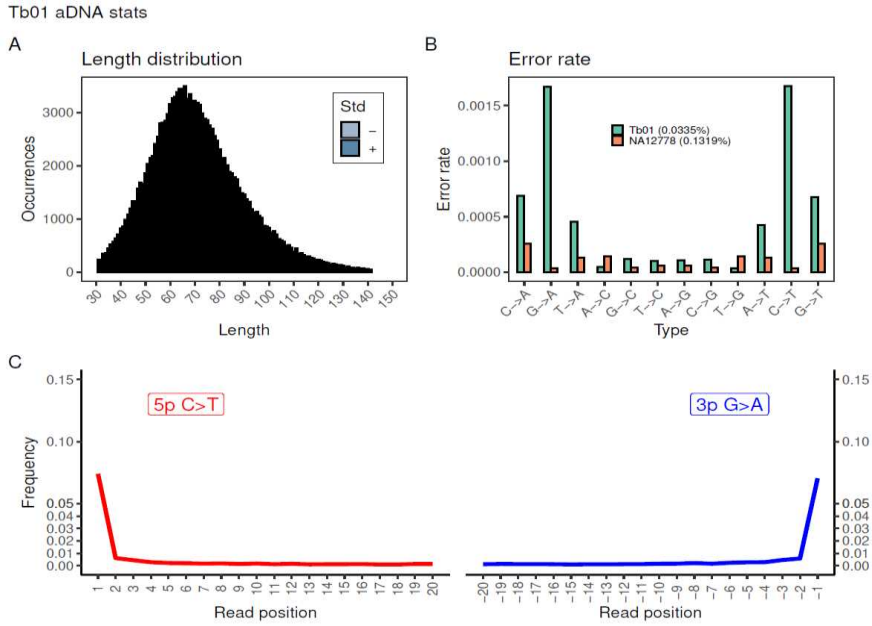


Figure 9.3. *Ancient DNA data validation on Tb01. (A) Read length distribution; (B) Error rate estimated using the chimpanzee genome as an outgroup and the modern genome NA12778 (from the 1000 Genomes Project) as an error-free individual and comparison with a down-sampled version of the NA12778 data (orange); (C) Post-mortem damage pattern.*

All samples were initially extracted with the WTR approach and sequenced multiple times. The main results of the sequencing runs, which were analysed with a pipeline developed in our laboratory, are shown in *Tables 9.1 and 9.2.*

Table 9.1. *Sequencing results on ancient Medieval individuals from Sant'Anna di Sopramonte.*

Sample ID	Indexes	Read leng. avg	% duplicates	% endogen. DNA	Avg depth (entire genome)	Sex ry	Avg depth MT	mtDNA Hg
Tb01	MPI101-LP6	56	2.44	5.13	0.02	M?	1.77	H5a+152
	MPI101-LP6	71	2.37	1.90	0.003	M	0.24	NA
	MPI99-LP1	66	1.57	0.53	0.001	F	0.52	NA
	MPI99-LP1	58	2.82	2.83	0.003	M?	0.37	NA
	MPI101-LP6	54	10.23	4.78	0.19	M	13.42	H5a2
	MPI101-LP06	55	18.00	4.59	0.09	M	8.07	H5a2
	MPI101-LP6	55	19.94	4.53	0.09	M	7.55	H5a2
Tb10	MPI101-LP3	84	4.71	8.26	0.02	F	6.92	U4a1b2
	MPI101-LP3	68	24.88	6.35	0.14	F	34.49	U4a1b2
	MPI101-LP03	79	24.19	4.92	0.09	F	31.62	U4a1b2
	MPI101-LP3	80	31.02	4.55	0.13	F	45.13	U4a1b2
Tb11	MPI98-LP1	79	0.12	1.33	1.33	F	1.33	U4a1
	MPI98-LP1	86	2.76	1.65	1.65	F	1.65	U4a2a
	MPI98-LP1	67	6.00	1.51	1.51	F	1.51	U4a1d
	MPI98-LP1	93	25.72	0.78	0.78	F	0.78	U4a1d
Tb14	MPI101-LP4	66	1.58	0.53	0.001	F	0.52	NA
	MPI101-LP5	73	2.38	0.54	0.001	F	1.87	H14a
	MPI115-LP4	61	13.12	0.42	0.003	F	4.73	H14a
	MPI115-LP4	68	32.87	0.35	0.006	F	8.88	H14a

Avg - average; Mt – mitochondrial; leng. – length; NA – not available, not possible to obtain data from the sample

Table 9.2. *Examples of contamination estimates (with confidence intervals).*

Sample ID	ContamMix	Schmutzi	HapCon
Tb01	0.962 (0.929-0.981)	0.01(0-0.02)	0.004 (0.001-0.008)
Tb10	0.993 (0.983-0.997)	0.01(0-0.02)	NA
Tb11	0.998 (0.988-0.999)	0.01(0-0.02)	NA
Tb14	0.994 (0.965-0.999)	0.01(0-0.02)	NA

NA means undetermined

Taking into account the low endogenous content obtained (<9%), we also tested the alternative MDE extraction approach on the sample Tb14. Table 9.3 summarises the extraction and library results on two remains, Tb14 and a sample from another archaeological site ("Palafitte" di Ledro, Trento), both extracted with both WTR and MDE methods.

Table 9.3. *Comparison of results obtained by using different extraction protocols.*

Sample ID	Extracted DNA QubitHS (ng/ul)		PostPCR Libraries QubitHS (ng/ul)		% Endogenous DNA	
	MDE	WTR	MDE	WTR	MDE	WTR
Tb14	1.17	0.66	16.70	11.50	0.54	0.52
Led6	1.05	0.82	10.50	12.00	2.13	0.98

MDE - Minimally Destructive Extraction; WTR - Whole Tooth Root protocol

The MDE protocol shows higher DNA concentrations and library yields for both samples, although the estimate of endogenous content after sequencing remained almost similar for Tb14.

The very few genomic reads obtained allowed for sex determination, confirming that the main individual buried in the "Tomb 1" could be a monk, while the others are likely nuns. We also obtained sufficient mtDNA data for a phylogenetic classification of all individuals (Table 9.1). The three different haplogroups identified suggest that the religious individuals buried in the cemetery probably had different origins, as these lineages are common in distinct geographical areas: from Northern and Eastern Europe (U4a1) to the Caucasus (H14a) and the Middle East (H5a). Further analyses of these ancient genomes are underway to clarify the genetic origins of the ancient individuals recovered from the archaeological site of Sant'Anna di Sopramonte.

10. References

Achilli A, Olivieri A, Semino O, Torroni A. Ancient human genomes—keys to understanding our past. Science. 2018;360:964–5. doi: 10.1126/science.aat725.

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. Proc Natl Acad Sci U S A. 2013;110:14308–13. https://doi.org/10.1073/pnas.1306290110.

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. Am J Hum Genet. 2007;80(4):759–68. doi:10.1086/512822.

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet. 2004;75(5):910–918. doi:10.1086/425590.

Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, et al. Saami and Berbers—An unexpected mitochondrial DNA link. Am J Hum Genet. 2005;76(5):883–886. doi: 10.1086/430073

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64. doi: 10.1101/gr.094052.109.

Allentoft M, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522(7555):167-172. https://doi.org/10.1038/nature14507.

Amorim A, Fernandes T, Taveira N. Mitochondrial DNA in human identification: a review. PeerJ. 2019;7:e7314. doi: 10.7717/peerj.7314.

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. doi: 10.1038/290457a0.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999;23:147. https://doi.org/10.1038/13779.

Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. 2012. control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Aneli S, Saupe T, Montinaro F, et al. The genetic origin of Daunians and the Pan-Mediterranean southern Italian Iron Age context. bioRxiv. 2021. doi: 10.1101/2021.07.30.454498.

Anthony DW. The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton (NJ): Princeton University Press; 2007.

Applebaum A. Red Famine: Stalin's War on Ukraine. New York: Doubleday; 2017.

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). Mol Ecol. 2000;9:1783–92. doi: 10.1046/j.1365-294x.2000.01068.x.

Abdulaeva G. Krims'ki tatary. Vid etnohenezu do derzhavnosti [Crimean Tatars. From Ethnogenesis to Statehood]. Kyiv: Zelenyi Pes; 2021. 406 p.

Balding DJ, Moltke I, Marioni J. Handbook of statistical genomics. 2019.

Bandelt H-J, Kong Q-P, Richards M, Macaulay V. Estimation of mutation rates and coalescence times: some caveats. In: Human mitochondrial DNA and the evolution of Homo sapiens. Nucleic Acids Mol Biol. 2006;18:47–90. doi: 10.1007/3-540-31789-9_4.

Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, et al. A 'Copernican' reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 2012;90(4):675-684. doi: 10.1016/j.ajhg.2012.03.002.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 2016;32:2817–23. doi: 10.1093/bioinformatics/btw327.

Bennett EA, Parasayan O, Prat S, Péan S, Crépin L, Yanevich A, Grange T, Geigl EM. Genome sequences of 36,000- to 37,000-year-old modern humans at Buran-Kaya III in Crimea. Nat Ecol Evol. 2023;7(12):2160-2172. doi:10.1038/s41559-023-02211-9.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9. doi: 10.1038/nature07517.

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. doi: 10.1126/science.aba9572.

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. doi: 10.1126/science.aay501.

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 Comput Biol. 2019;15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E, et al. The Paleo-Indian entry into South America according to mitogenomes. Mol Biol Evol. 2018;35:299–311. doi: 10.1093/molbev/msx267.

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. Int J Legal Med. 2004;118:294–306. doi: 10.1007/s00414-004-0466-z.

Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A. 1979;76:1967–71. https://doi.org/10.1073/pnas.76.4.1967.

Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA. Annu Rev Genomics Hum Genet. 2003;4:119-141. doi: 10.1146/annurev.genom.4.070802.110352.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185:3426-3440.e19. doi: 10.1016/j.cell.2022.08.004.

Capodiferro MR, Aram B, Raveane A, Migliore NR, Colombo G, Ongaro L, et al. Archaeogenomic distinctiveness of the Isthmo-Colombian area. Cell. 2021;184:1706-1723.e24. https://doi.org/10.1016/j.cell.2021.02.040

Cardinali I, Bodner M, Capodiferro MR, Amory C, Rambaldi Migliore N, Gomez EJ, et al. Mitochondrial DNA Footprints from Western Eurasia in Modern Mongolia. Front Genet. 2022;12:819337. doi: 10.3389/fgene.2021.819337.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: Whole-Genome Capture for the Targeted

Enrichment of Ancient DNA Sequencing Libraries. Am J Hum Genet. 2013;93(5):852-864. doi: 10.1016/j.ajhg.2013.10.002.

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003;33(S3):266-75. https://doi.org/10.1038/ng1113.

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 Res. 2012;22:821–6. doi: 10.1101/gr.134452.111.

Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;s13742-015-0047-8. doi: 10.1186/s13742-015-0047-8.

Chaubey G, Karmin M, Metspalu E, Metspalu M, Selvi-Rani D, Singh VK, et al. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evol Biol. 2008;8:227. https://doi.org/10.1186/1471-2148-8-227.

Chen YS, 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. Am J Hum Genet. 1995;57:133.

Clemente F, Unterländer M, Dolgova O, Amorim CEG, Coroado-Santos F, Neuenschwander S, Ganiatsou E, et al. The genomic history of the Aegean palatial civilizations. Cell. 2021;184(10):2565-2586.e21. doi:10.1016/j.cell.2021.03.039.

Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;5:a012567.

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. Proc Natl Acad Sci U S A. 2021;118. doi: 10.1073/pnas.2100901118.

Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. Improving access to endogenous DNA in ancient bones and teeth. Sci Rep. 2015;5(1):11184. https://doi.org/10.1038/srep11184

Duewer DL, Kline MC, Romsos EL, Toman B. Evaluating droplet digital PCR for the quantification of human genomic DNA: converting copies per nanoliter to nanograms nuclear DNA per microliter. Anal Bioanal Chem. 2018;410:2879–87. doi: 10.1007/s00216-018-0982-1.

Dür A, Huber N, Parson W. Fine-Tuning Phylogenetic Alignment and Haplogrouping of mtDNA Sequences. Int J Mol Sci. 2021;22:5747. https://doi.org/10.3390/ijms22115747.

Erkin Alaçamlı, Thijessen Naidoo, Şevval Aktürk, Merve N. Güler, Igor Mapelli, Kıvılcım Başak Vural, Mehmet Somel, Helena Malmström, Torsten Günther. READv2: Advanced and user-friendly detection of biological relatedness in archaeogenomics, BioRxiv 2024.01.23.576660; doi: https://doi.org/10.1101/2024.01.23.576660.

Ewens WJ. Mathematical population genetics. Springer; 2004.

Fisher RA. XXI.—on the dominance ratio. Proc R Soc Edinb. 1923;42:321-341.

Frank A, Lobry J. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene. 1999;238:65–77. https://doi.org/10.1016/S0378-1119(99)00297-8.

Frantz LA, Bradley DG, Larson G, Orlando L. Animal domestication in the era of ancient genomics. Nat Rev Genet. 2020;21:449–60. https://doi.org/10.1038/s41576-020-0225-0.

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. https://doi.org/10.1038/nature13810.

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. Proc Natl Acad Sci U S A. 2013;110:2223–7. https://doi.org/10.1073/pnas.1221359110.

Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Curr Biol. 2013;23:553–9. doi: 10.1016/j.cub.2013.02.044.

Garcia-Erill G, Albrechtsen A. Evaluation of model fit of inferred admixture proportions. Mol Ecol Resour. 2020;20(4):936–49. doi:10.1111/1755-0998.13171.

Gasparre G, Porcelli AM. The human mitochondrial genome: From basic biology to disease. Academic Press; 2020. 596 p.

Gelabert P, Schmidt RW, Fernandes DM, et al. Genomes from Verteba cave suggest diversity within the Trypillians in Ukraine. Sci Rep. 2022;12:7242. doi:10.1038/s41598-022-11117-8

Gerling C. Prehistoric mobility and diet in the West Eurasian steppes 3500 to 300 BC: an isotopic approach. Berlin, Boston: De Gruyter; 2015. 402 p. https://doi.org/10.1515/9783110311211.

Gnecchi-Ruscone GA, Khussainova E, Kahbatkyzy N, Musralina L, Spyrou MA, Bianco RA, et al. Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. Sci Adv. 2021;7:eabe4414. doi:10.1126/sciadv.abe4414.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.

Gopalan S, et al. Human genetic admixture through the lens of population genomics. Philos Trans R Soc Lond B Biol Sci. 2022;377(1852):20200410. https://doi.org/10.1098/rstb.2020.0410.

Govedarica VB, et al. Der Grabhügel 'Tarasova Mogila' bei der Stadt Orechov. 2006.

Graves JAM. Sex chromosome specialization and degeneration in mammals. Cell. 2006;124(5):901-914. doi: 10.1016/j.cell.2006.02.024.

Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science. 1999;283(5407):1476-1481.

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. doi: 10.1126/science.1188021.

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PLF, Uhler C, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008;134:416–26. doi: 10.1016/j.cell.2008.06.021.

Haak W, Lazaridis I, Patterson N, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11. https://doi.org/10.1038/nature14317.

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 Res. 2013;42:1111–6. doi: 10.1093/nar/gkt969.

Harney É, Cheronet O, Fernandes DM, Sirak K, Mah M, Bernardos R, et al. A minimally destructive protocol for DNA extraction from ancient teeth. Genome Res. 2021;31:472–83. doi: 10.1101/gr.267534.120.

Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107:1–8. doi: 10.1016/j.ygeno.2015.11.003.

Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet. 2010;19:R221–6. doi: 10.1093/hmg/ddq403.

Helgason A, Einarsson A, Guðmundsdóttir V, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, et al. The Y-chromosome point mutation rate in humans. Nat Genet. 2015;47:453–7. doi: 10.1038/ng.3171.

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312(5991):282–4. doi: 10.1038/312282a0.

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, et al. A mitochondrial stratigraphy for island southeast Asia. Am J Hum Genet. 2007;80:29–43. doi: 10.1086/510412.

Ho S. The molecular clock and estimating species divergence. Nat Educ. 2008;1(1):1-2.

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-del-Molino D, van Dorp L, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc Natl Acad Sci U S A. 2016;113(25):6886–6891. doi: 10.1073/pnas.1523951113.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet. 2001;2:353–66. doi: 10.1038/35072071.

Huang Y, Ringbauer H. hapCon: estimating contamination of ancient genomes by copying from reference haplotypes. Bioinformatics. 2022 Aug 2;38(15):3768-3777. doi: 10.1093/bioinformatics/btac390.

Hutchison CA, Newbold JE, Potter SS, Edgell MH. Maternal inheritance of mammalian mitochondrial DNA. Nature. 1974;251:536–8. https://doi.org/10.1038/251536a0.

Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet. 2009;124(6):579-591. doi:10.1007/s00439-008-0593-6.

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217–22. doi: 10.1038/ng1001-217.

Jobling MA, Hurles M, Tyler-Smith C. Human evolutionary genetics: Origins, peoples and disease. 2nd ed. New York: Taylor & Francis Inc.; 2014.

John Eid, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;323:133-138. doi:10.1126/science.1162986.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29:1682–4. doi: 10.1093/bioinformatics/btt193.

Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015;25:918–25. doi: 10.1101/gr.176552.114.

Juras A, Krzewińska M, Nikitin A, et al. Diverse origin of mitochondrial lineages in Iron Age Black Sea Scythians. Sci Rep. 2017;7:43950. doi:10.1038/srep43950

Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66. doi: 10.1093/molbev/mst010.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. doi: 10.1093/molbev/mst010.

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. Nat Genet. 2007;39:1251–5. doi: 10.1038/ng2116.

Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-626. https://doi.org/10.1038/217624a0.

Kingman JFC. The coalescent. Stoch Process Their Appl. 1982;13(3):235-248.

Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012;40:e3. doi: 10.1093/nar/gkr771.

Kivisild T. Maternal ancestry and population history from whole mitochondrial genomes. Investig Genet. 2015;6:1–10. doi: 10.1186/s13323-015-0022-2.

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. doi: 10.1126/science.abi5658.

Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15:356. doi: 10.1186/s12859-014-0356-4.

Krebs JE, Goldstein ES, Kilpatrick ST. Lewin's Genes XII. Burlington, MA: Jones & Bartlett Learning; 2018.

Kubicek P. The history of Ukraine. Westport (CT): Greenwood Press; 2008.

Kushniarevich A, Utevska O, Chuhryaeva M, et al. Genetic heritage of the Balto-Slavic speaking populations: A synthesis of autosomal, mitochondrial and Y-chromosomal data. PLoS One. 2015;10:e0135820. doi:10.1371/journal.pone.0135820.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015;3:1-8. doi:10.1016/j.bdq.2015.02.001.

Lazaridis I, Alpaslan-Roodenberg S, Acar A, Açıkkol A, Agelarakis A, et al. The genetic history of the Southern Arc: a bridge between West Asia and Europe. Science. 2022;26;377(6609):eabm4247 doi: 10.1126/science.abm4247

Lazaridis I. The evolutionary history of human populations in Europe. Curr Opin Genet Dev. 2018;53:21-27. doi: 10.1016/j.gde.2018.06.007.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536(7617):419-424. https://doi.org/10.1038/nature19310.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95. doi: 10.1093/bioinformatics/btp698.

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6. https://doi.org/10.1038/nature10231.

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. doi: 10.1093/bioinformatics/btp352.

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. https://doi.org/10.1038/s41586-021-04018-9.

Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 2012;5:337. doi: 10.1186/1756-0500-5-337.

Lindsay H, Kroman T, Kroman A. Methods in human skeletal biology. In: Bone and dental histology. 2013:361-395.

Lu Y, Shen Y, Wesley W, Ronald W. Next generation sequencing in aquatic models. In: Kulski JK, editor. Next generation sequencing-advances, applications and challenges. London: IntechOpen; 2016. p. 61–79.

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. doi: 10.1126/science.1109792.

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. https://doi.org/10.1038/nature18964.

Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, Patterson N, Reich D. The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes. bioRxiv. 2023. doi:10.1101/2023.04.06.535797.

Mallick S, Reich D. The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes. Harvard Dataverse. 2023. https://doi.org/10.7910/DVN/FFIDCW.

Malyarchuk BA, Derenko MV. Mitochondrial gene pool of UKRainians in the context of variability of whole mitogenomes in Slavic peoples. Genetika. 2023;59(1):106-114. doi:10.31857/S0016675823010083.

Malyarchuk BA, Derenko MV. Mitochondrial DNA variability in Russians and UKRainians: Implication to the origin of the Eastern Slavs. Ann Hum Genet. 2001;65:63-78. doi:10.1046/j.1469-1809.2001.6510063.x.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73. doi: 10.1093/bioinformatics/btq559.

Margulies M, Egholm M, Altman W, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80. doi:10.1038/nature03959.

Maricic T, Whitten M, Pääbo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE. 2010;5(11):e14004. doi: 10.1371/journal.pone.0014004.

Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. 2020;21:250. doi: 10.1186/s13059-020-02160-7.

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(7695):197-203. doi: 10.1038/nature25778.

Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, et al. The structure of human mitochondrial DNA variation. J Mol Evol. 1991;33:543–55. doi: 10.1007/BF02102807.

Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11:31–46. https://doi.org/10.1038/nrg2626.

Mielnik-Sikorska M, Daca P, Malyarchuk B, Derenko M, Skonieczna K, Perkova M, et al. The history of Slavs inferred from complete mitochondrial genome sequences. PLoS ONE. 2013;8(1):e54360. doi:10.1371/journal.pone.0054360.

Mittnik A, Wang CC, Svoboda J, Krause J. A molecular approach to the sexing of the triple burial at the Upper Paleolithic site of Dolní Věstonice. PLoS ONE. 2016;11:e0163019. doi: 10.1371/journal.pone.0163019.

Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010;19:R131–6. doi: 10.1093/hmg/ddq400.

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. Sci Rep. 2020;10:10700. doi: 10.1038/s41598-020-67445-0.

Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. 1987;155:335–50. doi: 10.1016/0076-6879(87)55023-6.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51. https://doi.org/10.1038/nrg2986.

Nikitin AG, Kochkin IT, June CM, et al. Mitochondrial DNA sequence variation in the Boyko, Hutsul, and Lemko populations of the Carpathian highlands. Hum Biol. 2009;81:43–58. doi:10.3378/027.081.0104.

Nikitin A, Newton J, Potekhina IM. Mitochondrial haplogroup C in ancient mitochondrial DNA from Ukraine extends the presence of East Eurasian genetic lineages in Neolithic Central and Eastern Europe. J Hum Genet. 2012;57(10):610-2. doi:10.1038/jhg.2012.69

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. bioRxiv. 2021;2021.05.26.445798. doi: 10.1126/science.abj6987.

O'Rourke DH, Raff JA. The human genetic history of the Americas: the final frontier. Curr Biol. 2010;20:R202–7. doi: 10.1016/j.cub.2009.11.051.

Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4. doi: 10.1093/bioinformatics/btv566.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. Erratum: The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555(7697):190-196. https://doi.org/10.1038/nature25738.

Oleksyk TK, Wolfsberger WW, Weber AM, Shchubelka K, Oleksyk OT, Levchuk O, et al. Genome diversity in Ukraine. GigaScience. 2021;10(1):giaa159. doi:10.1093/gigascience/giaa159.

Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, et al. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. Curr Biol. 2016;26(2):270-275. doi: 10.1016/j.cub.2015.12.019.

Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, et al. Ancient DNA analysis. Nat Rev Methods Primers. 2021;1:1–26. https://doi.org/10.1038/s43586-020-00011-0.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:74–8. doi: 10.1038/nature12323.

Parzinger H. Die frühen Reiternomaden der eurasischen Steppe: Neue Lebens- und Gesellschaftsformen zwischen Jenissei und Unterer Donau. In: Seipel W, editor. Das Gold der Steppe. Fürstengräber jenseits des Alexanderreichs. Wien; 2009. p. 16–29.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93. doi: 10.1534/genetics.112.145037. Epub 2012 Sep 7.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190. https://doi.org/10.1371/journal.pgen.0020190.

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, et al. Genetic analyses from ancient DNA. Annu Rev Genet. 2004;38:645–79. doi: 10.1146/annurev.genet.37.110801.143214.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. Genome Biol. 2016;17:60. doi: 10.1186/s13059-016-0918-z.

Penske S, Rohrlach AB, Childebayeva A, et al. Early contact between late farming and pastoralist societies in southeastern Europe. Nature. 2023;620:358–365. doi:10.1038/s41586-023-06334-8.

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. doi: 10.1126/science.1141564.

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. https://doi.org/10.1371/journal.pone.0129102.

Pinhasi R, Fernandes DM, Sirak K, Cheronet O. Isolating the human cochlea to generate bone powder for ancient DNA analysis. Nat Protoc. 2019;14:1194–205. https://doi.org/10.1038/s41596-019-0137-7.

Plokhy S. The Gates of Europe: A History of Ukraine. New York: Basic Books; 2015.

Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. Curr Biol. 2016;26(6):827-833. doi: 10.1016/j.cub.2016.01.037.

Posth C, Yu H, Ghalichi A, Rougier H, Crevecoeur I, et al. Palaeogenomics of Upper Palaeolithic to Neolithic European hunter-gatherers. Nature. 2023;615:117–126. https://doi.org/10.1038/s41586-023-05726-0.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 2013;341(6145):562-565. doi: 10.1126/science.1237619.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59. doi: 10.1093/genetics/155.2.945.

Pritchard JK. An Owner's Guide to the Human Genome: An introduction to human population genetics, variation, and disease. Stanford University; 2023.

Pshenichnov A. Struktura genofonda Ukraintsev po dannym o polimorfizme mitokhondrialnoi i Y khromosomy: avtoref. dis. kand. biol. nauk: 03.00.15. Moskva; 2007. 30 p. [Russian]

Pshenichnov A, Balanovsky O, Utevska O, Metspalu E, Zaporozhchenko V, Agdzhoyan A, et al. Genetic affinities of UKRainians from the maternal perspective. Am J Phys Anthropol. 2013. doi:10.1002/ajpa.22371.

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. Am J Hum Genet. 2007;81:559–75. doi: 10.1086/519795.

Pustovalov S. Economy and social organization of northern Pontic steppe – forest-steppe pastoral populations: 2750–2000 BC (Catacomb Culture). Baltic-Pontic Stud. 1994;2:86–134.

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet. 1999;23:437–41. https://doi.org/10.1038/70550.

R Core Team. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.

Racimo F, Renaud G, Slatkin M. Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. PLoS Genet. 2016;12:e1005972. https://doi.org/10.1371/journal.pgen.1005972.

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. https://doi.org/10.1038/nature08835.

Rassamakin Y. Carpathian imports in the graves of the Yamnaya Culture on the Lower Dnieper. Some problems of chronology and connections in the Black Sea steppes during the Early Bronze Age. In: Biehl PF, Rassamakin YY, editors. Import and Imitation in Archaeology. 2008. p. 51-87.

Rassamakin Y. The Eneolithic of the Black Sea steppe: dynamics of cultural and economic development 4500–2300 BC. In: Late prehistoric exploitation of the Eurasian steppe. 1999. p. 59-182.

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. Proc Natl Acad Sci U S A. 2014;111:15474–9. doi: 10.1073/pnas.1409328111.

Reich D et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053–60. https://doi.org/10.1038/nature09710.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012;488:370–4. https://doi.org/10.1038/nature11258.

Renaud G, Slon V, Duggan AT, Kelso J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 2015;16:224. https://doi.org/10.1186/s13059-015-0776-0.

Richards MB, Soares P, Torroni A. Palaeogenomics: mitogenomes and migrations in Europe's past. Curr Biol. 2016;26:R243–R246. https://doi.org/10.1016/j.cub.2016.01.044.

Reitsema LJ, Mittnik A, Kyle B, Catalano G, Fabbri PF, Kazmi AC, et al. The diverse genetic origins of a Classical period Greek army.

Proc Natl Acad Sci U S A. 2022;119(41):e2205272119. https://doi.org/10.1073/pnas.2205272119.

Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol Sci. 2015;370:20130624. doi: 10.1098/rstb.2013.0624.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002;298:2381-5. doi:10.1126/science.1078311.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51. https://doi.org/10.1186/gb-2013-14-5-r51.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7. doi: 10.1073/pnas.74.12.5463.

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. doi: 10.1371/journal.pone.0034131.

Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13:745–53. doi: 10.1038/nrg3295.

Schmidt RW, Wakabayashi K, Waku D, Gakuhari T, Koganebuchi K, Ogawa M, Karsten JK, Sokhatsky M, Oota H. Analysis of ancient human mitochondrial DNA from Verteba Cave, Ukraine: insights into the Late Neolithic-Chalcolithic Cucuteni–Tripolye culture. Anthropol Sci. 2020;128(1):1-10. doi:10.1537/ase.200205

Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012;13(12):878-890. doi: 10.1038/nrg3275.

Schönherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res. 2023;51:W263–8. doi: 10.1093/nar/gkad284.

Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KAS, Willerslev E, et al. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 2012;13:178. https://doi.org/10.1186/1471-2164-13-178.

Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88. https://doi.org/10.1186/s13104-016-1900-2.

Schönherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res. 2023;51:W263–W268. doi:10.1093/nar/gkad284.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, et al. The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. Science. 2000;290:1155-1159. doi:10.1126/science.290.5494.1155.

Sharma A, Jaloree S, Thakur RS. Review of clustering methods: toward phylogenetic tree constructions. In: Proceedings of International Conference on Recent Advancement on Computer and Communication. Singapore: Springer; 2018. p. 475-480.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550:345–53. https://doi.org/10.1038/nature24286.

Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45. doi:10.1038/nbt1486.

Shishlina N, et al. Paleoecology, subsistence, and 14 C chronology of the Eurasian Caspian Steppe Bronze Age. Radiocarbon. 2009;51(2):481–499. doi:10.1017/S0033822200055879.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423(6942):825-837. https://doi.org/10.1038/nature01722.

Skoglund P, Storå J, Götherström A, Jakobsson M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. J Archaeol Sci. 2013;40:4477–82. https://doi.org/10.1016/j.jas.2013.07.004.

Slimak L, Zanolli C, Higham T, Frouin M, Schwenninger JL, Arnold LJ, Demuro M, et al. Modern human incursion into Neanderthal territories 54,000 years ago at Mandrin, France. Sci Adv. 2022;8. doi:10.1126/sciadv.abj9496.

Slon V, Viola B, Renaud G, Gansauge M-T, Benazzi S, Sawyer S, et al. A fourth Denisovan individual. Sci Adv. 2017;3:e1700186. doi: 10.1126/sciadv.1700186.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H-J, et al. The archaeogenetics of Europe. Curr Biol. 2010;20:R174–83. doi: 10.1016/j.cub.2009.11.054.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009;84:740–59. doi: 10.1016/j.ajhg.2009.05.001.

Roostalu U, Kutuev I, Loogväli E-L, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol. 2007;24(2):436–448. doi:10.1093/molbev/msl173.

Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. Nat Rev Genet. 2019;20:323–40. https://doi.org/10.1038/s41576-019-0119-1.

State Committee of Statistics of Ukraine. Державний комітет статистики України. http://2001.UKRcensus.gov.ua/results/general/nationality/donetsk/

Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Curr Biol. 2010;20:R188–93. https://doi.org/10.1016/j.cub.2009.11.052.

Sulimirski T. The Scyths. In: The Cambridge History of Iran. 1995. p. 149-199. doi:10.1017/CHOL9780521200912.005.

Telegin DY, Pustovalov SZ, Kovalyukh NN. Relative and absolute chronology of Yamnaya and Catacomb monuments the issue of co-existence. In: Chernyakov IT, Kaiser E, Klochko VI, KoSko A, Kovalyukh NN, Kruts VA, et al. The foundations of radiocarbon chronology of cultures between the Vistula and Dnieper: 3150-1850 BC. Adam Mickiewicz University, Eastern Inst., Inst. of Prehistory; 1999.

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

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. Harvesting the fruit of the human mtDNA tree. Trends Genet. 2006;22:339–345. doi: 10.1016/j.tig.2006.04.001.

Torroni A, Lott MT, Cabell MF, Chen YS, 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. Am J Hum Genet. 1994;55:760.

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. Am J Phys Anthropol. 1994;93:189–99.

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. Am J Hum Genet. 1993;53:563.

Tuross N, Campana MG. Ancient DNA. In: The Science of Roman History: Biology, Climate, and the Future of the Past. Princeton University Press; 2018. https://doi.org/10.2307/j.ctvc772w1.

Underhill PA, Kivisild T. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007;41:539–64. doi: 10.1146/annurev.genet.41.110306.130407.

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. https://doi.org/10.1038/s41586-021-03224-9.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34:666–81. doi: 10.1016/j.tig.2018.05.008.

Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mut. 2009;30:E386–94. doi: 10.1002/humu.20921.

Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, et al. High-throughput DNA sequencing of ancient wood. Mol Ecol. 2018;27:1138–54. doi: 10.1111/mec.14514.

Watson JD. Molecular biology of the gene. 7th ed. Benjamin-Cummings Publishing Company; 2014.

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. R Soc Open Sci. 2016;3:160239. https://doi.org/10.1098/rsos.160239.

White TD, Black MT, Folkens PA. Human osteology. Academic Press; 2011.

Wilkins JF. Unraveling male and female histories from human genetic data. Curr Opin Genet Dev. 2006;16:611–7. doi: 10.1016/j.gde.2006.10.004.

Wright S. Evolution in mendelian populations. Genetics. 1931;16(2):97.

Yakubova L. Etnonatsionalna istoriia Donbasu: tendentsii, superechnosti, perspektyvy v svitli suchasnoho etapu Ukrainskoho natsiotvorennia. Kyiv: In-t istorii Ukrainy NAN Ukrainy; 2014. 108 p. [Ukrainian].

Yan C, Duanmu X, Zeng L, Liu B, Song Z. Mitochondrial DNA: distribution, mutations, and elimination. Cells. 2019;8(4):379. doi: 10.3390/cells8040379.

Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res. 2016 Jul 8;44(W1):W64-9. doi: 10.1093/nar/gkw247.

Wickham H. ggplot2: elegant graphics for data analysis. Springer-Verlag New York; 2016.

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. Front Bioeng Biotechnol. 2020;8:1032. https://doi.org/10.3389/fbioe.2020.01032.

Zickler D, Kleckner N. Recombination, pairing, and synapsis of homologs during meiosis. Cold Spring Harb Perspect Biol. 2015;7(6):a016626. doi: 10.1101/cshperspect.a016626.

Антоненко БА, Пиоро ИС, Самойленко ЛГ. Отчет о работе Ворошиловградской археологической экспедиции КГУ в 1987 г. Киев; 1988. Науковий архів Інституту археології НАНУ, фонд експедицій, спр. 1987/106. Машинопис рос. мовою; 95 с., іл.: 9 альбомів 730 рис.

Балановский ОП. Изменчивость генофонда в пространстве и времени: синтез данных о геногеографии митохондриальной ДНК и Y-хромосомы. Автореферат диссертации на исследование ученой доктора биологических наук. Москва; 2012.

Бондарь НН, Антоненко БА, Васильченко СА, Пиоро ИС, Самойленко ЛГ. Отчет о работе Каменской археологической экспедиции Киевского госуниверситета в 1976 г. Киев; 1977. Науковий архів Інституту археології НАНУ, фонд експедицій, спр. 1976/79. Машинопис рос. мовою; 48 с., іл.: альбом № 130 арк., альбом № 225 арк.

Бондарь НН, Антоненко БА, Васильченко СА, Пиоро ИС, Чмыхов НА. Отчет о работе Никопольской археологической экспедиции научно-исследовательского сектора Киевского госуниверситета в зоне сооружения II очереди Никопольского орошаемого массива в 1977 г.

Киев; 1978. Науковий архів Інституту археології НАНУ, фонд експедицій, спр. 1977/80. Машинопис рос. мовою; 109 с., іл.: альбом № 1 - 153 рис., альбом № 2 - 184 рис., альбом № 3 - 126 рис.

Самойленко ЛГ, Піоро ВІ. Знахідки давньогрецьких амфор в скіфських курганах Дніпропетровщини. Морська торгівля в Північному Причорномор'ї: збірка наук. статей. Київ; 2001. с. 26-42. Available from: http://vitaantiqua.org.ua/wp-content/uploads/2023/11/103_Pioro_Samoilenko.pdf

11. Acknowledgements

This PhD journey was challenging, but achieving my goal was possible thanks to the many wonderful people who supported me along the way. I'd like to express my appreciation for their support here, while saving a more personal acknowledgment for them individually.

I would like to acknowledge Professor **Alessandro Achilli** for his guidance and the academic opportunities he provided throughout my PhD research, especially granting me access to my dream workspace, the archaeogenetics lab. Professor, I also thank you for the openness and courage to welcome a foreigner into your team so readily.

I am grateful to all the Human and Animal Population Genomics Laboratory members, especially Professors **Antonio Torroni, Ornella Semino, Ana Olivieri, Luca Ferretti** and Drs. **Viola Grugni** and **Patrizia Chiari**. Many thanks to my colleagues and friends: **Elisabetta, Vincenzo, Giulia, and Damiano**.

I would like to express my gratitude to the current and former members of the Achilli Laboratory: **Nicola, Ana, Rosalinda, Giacomo, Ilaria, Anna, Gary, Aleksandro, Valeria, Vittoria, Serena, and Marco**. My PhD years were enjoyable thanks to our time together, both in academic and non-academic conversations, which I will miss.

Thank you, Drs. **Hovirag Lancioni**, and **Pavlo Shydlovskyi** for dedicating your time to reviewing this thesis. Special thanks to all the collaborators on the projects I developed or contributed to: **Svitlana Arbuzova, Liubov Samoilenko, Claudia Sharapova, Luca Pagani, Yevhen Zakharhcenko.** I am also grateful to all the donors whose DNA contributed to this work.

I would like to express my gratitude to Prof. **Lubov Atramentova**, Prof. **Valeriy Myasoedov**, and Dr. **Irina Meshcheryakova** for providing me with a solid foundation of knowledge and valuable experience.

I am deeply grateful to my family for their unwavering support, for standing by my decisions, and for granting me the freedom to determine my own path. To my mother, **Olga**, thank you for your incredible courage and warmth in embracing everything I do, no matter how far it takes me. Я люблю тебе, **матусю**, сильно-сильно! Дякую за твою підтримку у всьому, що я тільки не вигадаю! To my father, **Olexandr**, thank you for instilling in me a deep pride in being Ukrainian and for serving as an example of strength and bravery in the face of immense challenges. Your dedication to protecting our country, now in its third year of enduring a full-scale war, shows me that nothing is impossible when faced with courage. Дякую, **тату**, за наш захист і сміливість, за те, що завдяки таким людям як ти, я можу гордо називати себе українкою і мені є куди повертатися додому. Я люблю тебе!

My heartfelt thanks go to my friends, both old and new. Without you, facing the challenges of the past three years would have been far more difficult! Your friendship has been one of life's greatest gifts to me. **Mariia**, thank you for always being there, no matter the thousands of kilometres between us. You've shared all my emotions and kept me going with renewed energy. Люблю тебе, моя краща подруго!

Ana, I was incredibly lucky to meet you here. Despite the relatively short time we've known each other, you've become a true friend. I am grateful for the joyful moments we shared and your support, especially at the beginning of full-scale war in Ukraine, even though you did not know me well that time, you were present for me since then. Te quiero, Banana. Many thanks to **Elisabetta** and **Vincenzo** for the fun and jokes we shared, as well as your support and understanding. **Rosalinda**, thank you for your warmth and kindness; your presence in the lab always brought brightness. And **Mico**, thank you for being a good neighbour and friend in the collegio, I could always rely on you.

Even though we are far apart, **Olha, Valeriy, Anastasiia, Roman, Victoriia**, and **Inna** have remained close in my heart. Thank you all for years of friendship and love.

To my dear husband, **Giuseppe**, my deepest thanks for your unwavering support, faith, and love. Thank you for being there through every moment, from the darkest times to the happiest. I think you believe in me more than I believe in myself, never doubting my choices and always ready to embrace my craziest ideas. I am truly grateful for your constant encouragement, mental support and for being my personal shield against life's challenges. Ti amo! I also want to extend a special thanks to your family for their genuine care and kindness.